Practical applications of text analytics for Serious Mental Illness

Jackson, Richard George

Awarding institution:
King's College London

King's College London

Thesis submitted for the degree of Doctor of Philosophy

Practical Applications of Text Analytics for Serious Mental Illness

Richard Jackson

supervised by

Prof. Robert Stewart *(prim.)*

Prof. Richard Dobson *(sec.)*

March 23, 2019

Abstract

This thesis is an exploration of a problem that exists between cutting edge Natural Language Processing (NLP) methodologies and their real world exploitation in clinical research. I detail the development and validation of a range of NLP methodologies on clinical records, with a specific focus on the case of the symptomatology of Serious Mental Illness (SMI). This publication based thesis covers five main themes:

- Pre-work to describe the field of NLP within the context of clinical data
- The proposition, development and evaluation of the TextHunter desktop application, a suite of high-throughput tools to overcome bottlenecks in the development of NLP applications
- The application of the tools to the novel domain of SMI symptomatology, enabling the development of language models for 46 symptom concepts with a median F1 score of 0.87, and enabling the profiling of symptom distribution amongst 7 962 patients, based on discharge summaries
- A knowledge discovery project using artificial neural networks and clustering techniques, to identify real world patterns of symptom depiction in clinical free text. Here, I demonstrate a granularity and diversity of vocabulary beyond what is described in standard clinical terminologies.
- A commentary on the realities of text analytics in the NHS, and the development of a software architecture 'CogStack' to address these. This culminated in the establishment of the Clinical Analytics Platform at King's College Hospital.

Word Count: 69 424

Contents

List of Figures

List of Tables

Declaration

Acknowledgements

To Rob Stewart, I would like to express my eternal gratitude for giving me the opportunity to complete a PhD. I am thankful for your unending patience, wisdom, far-sightedness, discretion and for the freedom to explore my ideas.

To Richard Dobson, thank you for your advice throughout and the opportunities you helped to create for me.

To Angus Roberts, an enormous thank you for your detailed technical review of my thesis and a plethora of helpful editorial comments.

To Clive Stringer, thank you for creating the opportunity for me to execute the CogStack project.

The following people have been especially helpful to me throughout the last four years: Richard Hayes, Johnny Downs, Cass Johnston, Genevieve Gorrell, Rashmi Patel, Steve Newhouse.

I would also like to say thanks to the following for their friendship and support:

Chin-Kuo Chang, Nishamali Jayatilleke, Amos Folarin, Anna Kolliakou, Michael Ball, Karolina Bogdanowicz, Guiliana Kadra, Hitesh Shetty, Ryan Little, Megan Pritchard, Matthew Broadbent, Robbie Mallah, Clare Taylor, Ismail Kartoglu, Honghan Wu, Alex Tulloch, Amelia Jewell, Andrea Newhouse, Ehtesham Iqbal, Max Kertz, Dan Leirer.

To my parents - thanks for everything.

To my wife Ellie and daughter Charlotte, all of my love, as always.

Introduction

The Future of Health

In 2014, NHS England released a document called the Five Year Forward View, outlining a bold vision to transform the NHS in England to address severe structural problems that will lead to a funding gap of £30 billion by 2020. Amongst the recommendations was the recognition of a world undergoing rapid technological change, leading to the aspiration of a fully paperless NHS in the UK by the year 2020 [2]. The implication for such an ambition may be a permanent change to the way that healthcare is provisioned in the UK. The benefits are cited as new opportunities for data sharing between patients and their doctors, greater efficiency of inter and intra-organisational data management, and perhaps most laudable of all, the possibilities of gaining greater insight into disease via harnessing masses of electronic data hitherto locked inside filing cabinets around the country. Ultimately, such operations seek to reduce the cost and increase the effectiveness of care. However, critics of the transformation argue that the digitalisation of health represents a fundamental change in the nature of the doctor/patient relationship, and responses to questions of ethics and information governance are insufficiently developed to cope with the rate of change. Nevertheless, with diverse stakeholder interests throughout the sector including clinicians, support staff, business intelligence, Trust executives, clinical and academic research leaders, central government policy makers and above all patients, large scale technological advancement in UK healthcare is all but inevitable.

The focus of this PhD is the application of informatics methods to address a small portion of the challenges and opportunities resulting from the changes the NHS is undergoing. Specifically, as the NHS moves to predominantly paperless information systems, my work explores how practical applications of techniques from the field of natural language processing (NLP) can be used to address real world problems in data exploitation, with case studies from a large mental health Trust, the South London and Maudsley NHS Foundation Trust, and a large acute care hospital, King's College Hospital Foundation Trust.

Research Questions

After covering the fundamentals of NLP and the history of clinical NLP in chapter 2, I attempt to develop an argument that the development of large clinical NLP frameworks are problematic and overly complicated for many use cases, and have created a divide between solving real world clinical data problems and clinical language as a field of study. I argue that minimal adaptation of existing algorithms and approaches (beyond the creation of training data specific to a task) are sufficient to produce good results in many clinical text classification problems, and that focussing on simplicity and practical concerns may provide a route to greater uptake of NLP in the UK clinical environment.

In the first half, the work concerns the application of Support Vector Machines to text classification of the symptomatology of serious mental illness (SMI). Here, I explore what results can be achieved from a relatively simple NLP pipeline, and how packaging this approach into a user orientated data collection, annotation, model construction and evaluation system might create access to some of the capabilities of NLP for users that would otherwise have none.

There is intuitive value of the development of the global clinical language resource, SNOMED-CT. The third quarter explores the use of word embeddings to identify important novel vocabulary of SMI that is in everyday usage in an NHS Trust, but exists outside of SNOMED-CT. The purpose of this chapter is an exploration of the disconnect between SNOMED-CT and real world clinical vocabulary, in an attempt to establish a methodology for building lexical resources suitable for use with the systems demonstrated in chapters 4 and 6 .

The final quarter asks the question: 'what happens if an NHS Trust I.T. function has access to a range of open source NLP software solutions with proven business value?'. Here, I focus on the opportunities created for NHS business analytics when an integrated information retrieval and extraction system is implemented in a large acute NHS Foundation Trust, using a combination of off-the-shelf open source software and an integration codebase designed with NHS systems in mind.

Objectives

The following objectives are defined for this PhD:

1. Develop background expertise via collaboration with external NLP research groups, and jointly investigate sentence classification approaches over a small selection of negative symptom concepts

2. Create software and methods to streamline the most practical sentence classification approach, in order that the solution might scale to many concepts

3. Validate the software over a large range of SMI symptomatology concepts, reporting findings and descriptive data of the resulting information

4. Investigate emerging NLP technologies to provide guidance for future research embarking on similar objectives

5. Ensure practical usage objectives are met by building an architecture capable of near real-time processing of new clinical data of symptom models and other NLP applications, considering information governance and data management/visualisation requirements

Novel contribution of this thesis to existing literature

- The validation of several pre-existing NLP applications in a variety of specific research contexts, directly contributing to results in several publications

- The development of the TextHunter program, an NLP suite that abstracts and simplifies many common NLP tasks into a single system, offering a degree of self provisioning of NLP analytics for non-expert users

- The scaling of the TextHunter program over a large selection of serious mental illness symptom concepts, producing novel symptomatology profiles over a large cohort of SMI sufferers

- An exploration of the value of artificial neural networks and clustering algorithms for exploring novel depictions of symptomatology in clinical free text

- The development of the CogStack, a software architecture that is capable of real time, large scale information extraction and retrieval, enabling robust, efficient use of many NLP applications with a low cost/management overhead

Publications (First Author)

- TextHunter - A User Friendly Tool for Extracting Generic Concepts from Free Text in Clinical Research [3]

- Natural Language Processing to Extract Symptoms of Severe Mental Illness from Clinical Text: The Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) Project [4]

- Knowledge Discovery for Deep Phenotyping Serious Mental Illness from Electronic Health Records [5]
- CogStack - Experiences of Deploying Integrated Information Retrieval and Extraction Services in a Large NHS Foundation Trust [6]

Publications (Co-Author)

- Clinical predictors of antipsychotic use in children and adolescents with autism spectrum disorders: a historical open cohort study using electronic health records [7]
- Novel psychoactive substances: An investigation of temporal trends in social media and electronic health records [8]
- Delays to diagnosis and treatment in patients presenting to mental health services with bipolar disorder [9]
- Association of cannabis use with hospital admission and antipsychotic treatment failure in first episode psychosis: an observational study [10]
- Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record derived data resource [1]
- Mood instability is a common feature of mental health disorders and is associated with poor clinical outcomes [11]
- The Effect of Clozapine on Premature Mortality: An Assessment of Clinical Monitoring and Other Potential Confounders [12]
- Negative symptoms in schizophrenia: a study in a large clinical sample of patients using a novel automated method [13]
- Identification of Adverse Drug Events from Free Text Electronic Patient Records and Information in a Large Mental Health Case Register [14]
- Extracting antipsychotic polypharmacy data from electronic health records: developing and evaluating a novel process [15]
- Delays before Diagnosis and Initiation of Treatment in Patients Presenting to Mental Health Services with Bipolar Disorder [16]
- Investigation of negative symptoms in schizophrenia with a machine learning text-mining approach [17]

- Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records [18]
- Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register [19]

Presentations/Workshops

- Clinical Text De-identification and other Large Scale Processing Tasks in Resource Constrained Environments, International Population Data Linkage Conference 2016
- EHR Free Text Anonymisation workshop, International Population Data Linkage Conference 2016
- Cognition-DNC - Making data available for clinical research The Farr Institute International Conference 2015
- TextHunter - A User Friendly Tool for Extracting Generic Concepts from Free Text in Clinical Research AMIA 2014 Annual Symposium (presenting [3])
- Finding Negative Symptoms of Schizophrenia in Patient Records, Recent Advances in Natural Language Processing 2013 (presenting [20])

Thesis Structure

Chapter 1 The thesis begins with a background exploration of the nature of EHRs, with a special reference to mental health. Here, I discuss the evolution of clinical informatics to address some of the challenges posed by electronic health data

Chapter 2 Following this background, I introduce generic NLP technical concepts which are used throughout the rest of the thesis. I put these concepts into context via a review of existing clinical NLP work, and the validation of some existing applications.

Chapter 3 In this chapter, I introduce an argument outlying some concerns in the current direction of clinical NLP research, and discuss the need for more rounded considerations of the clinical domain, in order for downstream processes to benefit from the labours of the academic community.

Chapter 4 Building on some preliminary work on concept extraction methods of the negative symptoms of schizophrenia, I introduce a machine learning sentence style classification method for information extraction. I discuss the creation and validation of the TextHunter system that streamlines this method to meet the needs of mental

health EHR epidemiology researchers. I describe the requirements for the application in detail, and provide an overview of the technical architecture to assist with its expansion and usage. I then discuss the CRIS-CODE project, which involves the application of TextHunter to a wide range of mental health concepts.

Chapter 5 Here, I explore the value of artificial neural networks and clustering to discover new symptom depictions in clinical text.

Chapter 6 In the final chapter, I turn my attention to practical matters of operating NLP systems at scale in production environments. I introduce the CogStack software.

Chapter 7 In my discussion, I review the overall theme of the thesis and explore the future role of NLP in the creation of an AI doctor.

Contribution Statement

Chapter 1 This introductory chapter is my own work

Chapter 2 This background chapter is my own work, with the following exceptions. The CRIS position paper referenced at the end of this chapter is the result of the work of many people. Full attributions can be found in the manuscript in Appendix A. Apart from TextHunter, the NLP applications described within it were developed by various individuals of the GATE team at the University of Sheffield. Apart from where specified, I organised, managed and contributed to the annotation effort to derive the classification performance statistics for the GATE NLP applications published in the paper. I authored the sections describing the GATE NLP work, with additional contributions regarding post-processing from Richard Hayes and Giuliana Kadra. References to using SAS for NLP were authored by Alex Tulloch

Chapter 3 This chapter is my own work

Chapter 4 The Negative Symptoms paper presented in this chapter was co-authored by Genevieve Gorrell, Angus Roberts, Robert Stewart and myself. I conceived the TextHunter concept in discussion with Robert Stewart and Richard Dobson. I wrote all of the TextHunter Code, and oversaw the annotation effort for the concepts described in the paper. Rashmi Patel, Robert Stewart, Michael Ball and Andrea Fernandez contributed to the annotation effort. Richard Hayes provided additional statistical support. The CRIS-CODE concept was conceived by Robert Stewart. I managed the annotation effort and provided TextHunter support for a team of

annotators comprising Rashmi Patel, Robert Stewart, Michael Ball, Anna Koulikou and Nishamali Jayatilleke, and conducted the subsequent analysis of the results

Chapter 5 I conceived the neural network and clustering concept for mental health symptomatology knowledge discovery. Clinical insight was provided in discussion with Robert Stewart and Rashmi Patel. Additional domain support was provided by Sumithra Velupillai and George Gkotsis

Chapter 6 The original de-identification Cognition algorithm was first proposed by Ismail Kartoglu. I reimplemented this into the form found within CogStack. Original and modified code attributions can be found within the source code. I developed the simulation code for clinical text mutation. Amos Folarin suggested the use of the Docker containerisation system. All other CogStack code until approximately mid Oct 2017 were developed by me. At this point, the project was open sourced, and is currently under community development

Chapter 7 This chapter is my own work.

Chapter 1

Background

1.1 Clinical Informatics Overview

Clinical informatics is a discipline on the interface between life science and computer science which has risen to prominence over the last two decades as a direct result of the expansion and maturation of electronic health data capture systems. The field originates from the need for more sophisticated tools and methods to respond to the deluge of clinical data that have arisen since Electronic Health Records (EHRs) have become mainstream in healthcare organisations. As a relatively new discipline, many definitions exist for the term "clinical informatics". However, the following from [21] serves as a useful guideline:

"*...[to] transform healthcare by analyzing, designing, implementing, and evaluating information and communication systems that enhance individual and population health outcomes, improve patient care, and strengthen the clinician-patient relationship.*"

The evolution of the field might be compared to the closely related field of bioinformatics, which has also enjoyed an elevation to prominence over the same period. Bioinformatics is predominantly concerned with the use of computation to facilitate understanding and analysis in molecular biology. While bioinformatics generally makes use of data generated via controlled experimentation in collaboration with geneticists, molecular biologists and other related fields, clinical informatics most often utilises secondary data captured from the EHRs, a generic term used to describe patient data captured by clinicians and administrative staff during their routine activities. Over the last two decades, both fields have sought to utilise computation to dramatically increase the quantity of data available for experimentation by making use of new technologies and algorithms, and specifically to clinical informatics, advances in Information Governance policy that enables secure access to key data without compromising patient anonymity.

In practical terms, clinical informatics is concerned with the development of methodologies to extract and clean data from large complex datasets; to design and make use of statistical methods to objectively study EHRs for both observational and predictive modelling; to develop applications and interfaces to serve actionable information to where it is needed most and to act as a conduit for the flow of knowledge between complex datasets and researchers with additional medical domain specific expertise, or systems that enable greater exploitation of raw clinical data. The cross disciplinary nature of the field necessitates that effective practice requires a diverse assortment of skills and experience, such as:

- Sufficient knowledge of the medical domain in which the individual is engaged
- Technical and programming skills to solve complex data issues
- Strong statistical and applied mathematics knowledge to be able to evaluate new developments and methodologies in the wider field of informatics and computer science
- Awareness of the wider business environment, and how changes in procedures can influence raw data collection
- Communication skills, to express complex issues succinctly to collaborators, without overlooking points of important detail

In addition, there may be requirements placed upon the informatician specific to the domain in which they are required, such as knowledge of parallel processing techniques and distributed computing to manage large datasets. Ultimately, the role is concerned with bridging the knowledge space between disparate fields of science, facilitating and directing collaboration amongst disease specialists to harness the power of high end computation.

In order to comprehend the role of clinical informatics in detail, it is necessary to understand something of the background of EHRs. The following sections provide an overview of the historical context of EHRs, the epidemiological beginnings of clinical informatics, through to the modern data science movement.

1.2 A Brief History of the Electronic Health Record and Mental Health

The ISO (International Organization for Standardization) definition of an EHR is [22]:

"... [a] repository of information regarding the health status of a subject of care, in computer processable form "

The ISO makes a slightly different definition for Electronic Medical Record (EMR), in that is should only concern medical information, although in practice the two terms are often used interchangeably. The EHR can be described as the digital equivalent of the paper based systems traditionally used to manage the medical information during the course of a patient's treatment by a clinical service. Historically, the concept of an EHR has been around for over 60 years. Although EHR systems are now available to all medical domains, curiously, some of the earliest references to electronic records originate in the field of mental health. One forward looking paper from 1962 describes the Maryland Psychiatric Case Register, a very early system that includes provisions for many common activities that EHRs are used for today, even highlighting the potential for secondary use in research - a concept which only recently has reached fruition [23]. Another EHR system, 'MSIS', was developed to meet a need for more efficient reporting systems for psychiatric care patients, which recognised privacy and confidentiality as key drivers in the evolution of such systems [24, 25]. The deep historical use of computerised systems in the provision of mental health care suggests an early requirement for efficient data management systems. Nevertheless, widespread usage of the EHR in mental health and indeed generally remained sparse until much later.

During the latter part of the 1990s, the personal computer had reached a general level of usability and a price point to meet the technical requirements that would support widespread use of EHRs throughout the medical community. However, many technical and conceptual issues would need to be resolved in order for their uptake to become more commonplace. Perhaps the most prominent issue to be identified in the early EHR era was how long-recognised conceptual differences in medical entities presented by different national and international institutions translate into EHR system design. For instance, traditional 'waterfall' style software development paradigms for information management systems of the time generally revolved around assumptions of static, complete data models in which all entities were known and standardised at the outset. Once the development of a system was complete, changes to it would be technically challenging if they violated the fundamental assumptions of the data model. The technical mindset of the era almost certainly would have struggled to keep pace with the clinical requirements around continuously evolving medical standards and practices [26]. Such issues have become recognised as one of the grand challenges in clinical data management and inter/intra-organisational data sharing; although existing initiatives such as SNOMED-CT [27], the HL7 messaging

system [28] and ICD [29] were devised to address issues of semantic and technological standardisation, the scale of the task has meant that implementing such initiatives have been only partially successful [30–33]. Indeed, even at a local level, Hicken et al. found that integrating legacy systems into a contemporary EHR data model posed substantial challenges in a large US healthcare organisation [34]. Nevertheless, the perceived benefits of conceptual EHR use from early adopters combined with shifting sociological perceptions of the role of technology in business evetually led to a deluge of EHR systems and providers. The uptake of EHRs by healthcare providers in western economies has risen sharply since 2000, mimicking the response to the digital age of many other domains heavily reliant on information systems. For instance, in primary care, usage of EHR systems in the USA doubled to 67.8% from 2005-2011, although significant variation in localities exist [35]. In the UK, uptake of EHRs in primary care is close to 100% [36] while a 2015 survey of 59 out of 235 acute, mental health and community care Trusts located throughout England found that 47 had either implemented, or were in the process of implementing an EHR system [37]. This is likely in response to policy development by the Department of Health [38] and increasing volumes of academic literature describing benefits as the main drivers. Specifically regarding mental health Trusts, EHR adoption may also be as high as 100% (Robert Stewart, personal communication).

Academic interest in the EHR has grown strongly year on year since around the mid 1990s. A simple search in PubMed for 'Electronic Health Record OR Electronic Medical Record OR EHR OR EMR' produces over 33 762 hits (August 2016) (Figure 1.1). A significant area of research in this early period concerned the impact of the EHR upon care at the point of delivery, privacy concerns, interoperability standards and cost/benefit tradeoff analyses for the early adopters [39–42], for instance, examining how the EHR affected communication efficiency between primary care and hospital pharmacy [43]. However, it was the establishment of the evidence-based medicine doctrine that would prove to have a profound effects on the research potential of the EHR.

1.2.1 Evidence-based medicine and EHR Epidemiology

Evidence-based medicine is an approach to medical decision making that has become standard practice in health doctrine throughout most of the world since the latter half of the 20th century. One definition of evidence-based medicine is [44]:

"*... the process of systematically reviewing, appraising and using clinical research findings to aid the delivery of optimum clinical care to patients* "

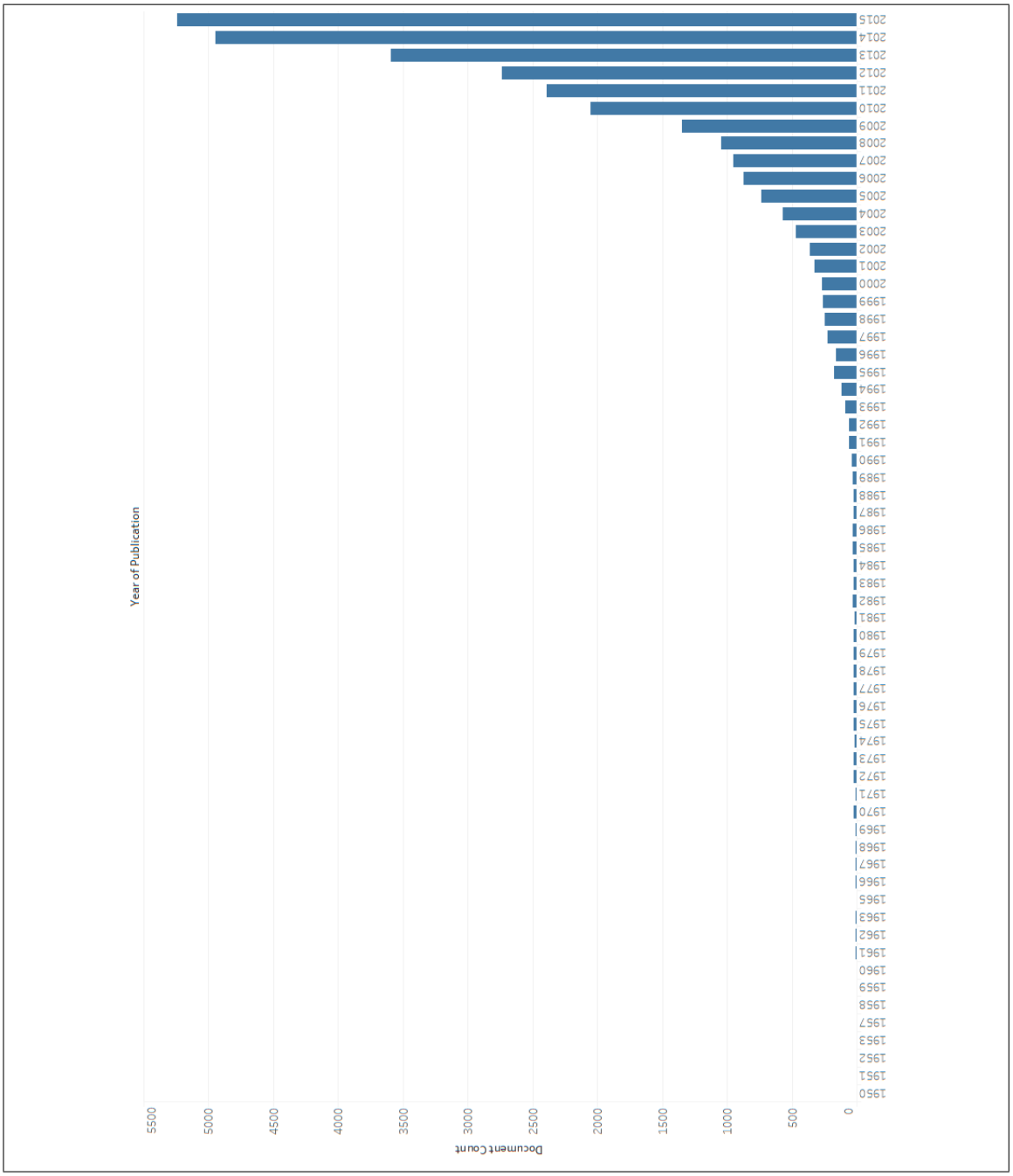


Figure 1.1: Year on year EHR articles in PubMed with hits on the search term 'Electronic Health Record OR Electronic Medical Record OR EHR OR EMR'

Evidence-based medicine originates from the growth of controlled trials as a means to assess the effectiveness of interventions in the 1940s [45]. However, it wasn't until the advocacy for the use of randomised controlled trials driven by Archie Cochrane, David Eddy, David Sackett, Gordon Guyatt [46–49] and others that the concept started to gain momentum as the international standard of developing medical practice. In 1993, the international Cochrane Collaboration was inaugurated, and today exists as one of the most prestigious sources of systemic reviews of medical literature and scientific knowledge upon which evidence-based medical principles are established.

Although the concept of evidence-based medicine is now ubiquitous, a humorous article examining other systems of medicine by Isaacs [50] serves as a reminder that evidence-based medicine only works when sufficient evidence is available. The gold standard of medical evidence, the randomised controlled trial (RCT), tends to be expensive in terms of cost and time to execute [51]. Therefore, evidence to support medical practice is generally limited relative to the number of potential clinical questions. To fuel the growing need for successful, informative RCTs, much of the preliminary scientific investment prior to the commencement of a trial concerns building a sufficiently large body of evidence to justify the investment. It was here that the role of the EHR as a source of observational research data emerged; both as a source of primary data collection at the point of care [52,53]; and the secondary reuse of clinical data [54] collected during routine administrative activities. Soon afterwards, some of the first studies using EHR data directly as raw data for research would appear [55, 56].

As far as the field of epidemiology was concerned, the advantages seemed obvious; the vast quantity of electronic data held within healthcare databases might be leveraged for observational studies that would otherwise be dependent on the expensive recruitment of cohorts to establish correlations between factors of interest [57]. Theoretically, data for such populations might be accessed via EHR systems when implemented within a sufficiently large healthcare organisation, or linked between disparate smaller organisations that shared compatible data models. With every passing day, the volume of data that might be made available to research would grow, and thus the older an EHR resource, the more valuable it would become. However, with the emergent opportunities of EHR research, new challenges would also appear, not least of which would be appropriate information governance.

1.3 'Big Data', Information Governance and the Safe Haven

From the year 2000 onwards, the growing interest in the secondary use of EHRs presented ethical and governance issues not often encountered in other forms of life science research. In many countries, data captured during the course of care is subject to strict laws regarding its access and use, in order to safeguard against fundamental ethical concerns and protect the relationship between doctors and patients. As increasing numbers of studies sought to use large numbers of individuals' private medical data in research, traditional models of directly obtaining consent became ineffectual. In order for EHR 'Big Data' to be made available for research, new models of information governance would be required to balance the public benefits against any harm that might result from undermining public confidence in healthcare provision. Developing appropriate national and trans-national policy for such requirements has proven to be far from trivial [58]. The high profile failure of the UK Care.Data programme [59] is illustrative of the delicate balance of trust that exists between researchers, data providers and the public. Care.Data was an attempt by NHS England to centralise and integrate health and social care data from all English NHS sites, creating one of the largest repositories of anonymous clinical information expressly intended for research use in the world. As details of the project were broadcast to the public domain, a significant backlash from both GPs and the public occurred over fears of lack of transparency and informed consent, re-identification concerns, hacking and access by private companies. In July 2016, NHS England announced the closure of the programme, citing lack of public confidence. Amongst the 'lessons learnt' from the exercise is that stakeholder engagement in such programmes, and the requirement to provision such records securely and anonymously is essential to their success [60].

Nevertheless, many EHR research programmes of varying scale exist throughout the UK. An increasingly popular method of managing information governance concerns is the use of the 'Safe Haven' model [61]. Although the term has no formal definition [62], it is generally interpreted as a secure location in which to conduct analysis upon EHR data, consisting of some or all of the following security elements:

Pseudonymisation Replacement of patient identifiers with a non-identifiable key

De-identification Removal of personal information from free text data

Access Audit Monitoring of user access logs to ensure data is used appropriately

Risk Assessment Data access approval decisions are often determined by committes of information governance specialists, patient interest group representatives, research leads and other relevant parties

Legal Arrangements Arrangement of data access agreements and enforcement of contractual obligations between data controllers and researchers regarding appropriate use

Information Toolkit Compliance with an appropriate level of the NHS Information Governance Toolkit, including data retention/destruction requirements

Public Engagement Activities to inform and reassure the public of the nature of EHR research activities

Training Education and training materials for researchers regarding the importance of information governance while working on EHR data

Specifically regarding mental health, a review by Stewart and Davis [63] surveying the international landscape of such resources identified 84 instances of projects that service data to organisations through a range of governance models. In the UK, there are several examples of safe havens capable of provisioning mental health EHR data. The largest include the Clinical Record Interactive Search (CRIS), described in detail in Chapter 3, operates a pseudonymised version of the South London and Maudsley NHS Trust EHR system, containing over 200 000 individual records of secondary mental heath [1, 64]; the Secure Anonymised Information Linkage Databank (SAIL) [65] offering project specific datasets from most of the NHS organisations in Wales; and the Mental Health Inpatient and Day Case dataset via the electronic Data Research and Innovation Service (eDRIS) in Scotland. Sustained successes in these ventures serve as evidence that viable solutions exist for the practical management of the principal concerns. However, it is worth noting that, to date, no major data breaches have been documented as a result of the activities of the safe haven model. Given the frequency of data breach in healthcare by other means [66], the seriousness of the consequences [67], and the protean nature of the EHR research dialogue, it is uncertain how public perceptions might change in such an event.

1.4 Realising the Potential of the EHR and Origins of Clinical Informatics

Early optimism for the potential for epidemiological research from EHRs was met with ambivalence or even misgivings amongst a sizeable group of academics, when the complexity and lack of precedent of dealing with data not specifically created for research emerged [57, 68–70]. While early papers acknowledge sources of bias, references to data quality indicators are rare. As largely exploratory work, this may be understandable due

to the simplistic nature of the variables used and the lack of precedent for managing quality issues appropriately, rather than adapting study designs to utilise secondary data. Specifically regarding mental health, a review of the issues by Munk-Jorgensen et al [71] raised the following concerns about secondary care clinical datasets collected via administrative means:

- The possibility for causative research is excluded due to the method of collection
- The data tends to be sparse, meaning highly granular analysis of specific variables can be problematic or impossible
- The distribution of data is uneven within cohorts. In secondary care case registrars, strong biases are exhibited according to the severity of the disease, since very ill patients are likely to present multiple times to the healthcare organisation. Conversely, less serious cases are likely to suffer from sparser data.
- In some conditions such as affective disorders, only a small percentage will be serious enough to reach secondary care. Therefore, any resulting research is unlikely to be of benefit to the majority of the affected patients.
- Predictive modelling of events that have a high value in preventative medicine may suffer if the event has a low frequency, due to the low predictive value of any predictors in a sparse dataset(for instance, the authors give the example of the rarity of suicide).

Given the temporal and geographical factors contributing to highly variable data (and in some cases, systemic weaknesses of the EHR research concept), the vision of a singular, granular, low noise, highly structured view of patient data has proven to be a long term aspiration rather than an immediate benefit of EHR adoption. Regardless, as EHRs become synonymous with the modern healthcare organisation, the emergence of tools and techniques for managing the issues associated with EHR data use might reasonably be ascribed to the field of 'Clinical Informatics'. Accordingly, several sub-specialisations of clinical informatics have arisen to address the issue of data quality.

1.4.1 Data linkage

The concept of data linkage has become a prominent way of combining disparate, complementary datasets to improve the quality and quantity of the available data. Although the theory of data linkage is trivial, in that datasets can be combined at an individual level on the basis of matching a primary key representing the individual in both datasets, in practice, such keys often do not exist in real world data [72]. This has given rise to an active

research area of best practice in data linkage, consisting of areas such as deterministic and probabilistic matching algorithms, data cleaning and standardisation, as well as raising a range of new questions about the potential ethical implications of constructing enormous datasets of largely unconsented research participants. Commonly selected datasets external to the EHR in the UK include the Hospital Episode Statistics dataset and the Office of National Statistics Mortality data.

Regarding clinical data in the UK, primary and secondary health services are generally delivered by distinct organisational units of the NHS, with little direct transfer of electronic information between them. Observing data in a single organisation therefore, is likely to only give a partial view of both individual's longitudinal records, and the population at large [73]. Here, data linkage between primary and secondary care units has been proposed as a method to fill in the gaps [74].

Specific to mental health, a common example of data linkage concerns the linking of psychiatric and criminal justice records. Via linked datasets, Wallace et al [75] and Fazel and Grann [76] observed an increased risk of offending amongst those with serious mental illness, identifying co-morbid substance abuse as a factor. Similarly, Herinckx et al identified that misdemeanants who had their case heard by a specialist mental health court were at substantially less risk of re-offending [77].

Today, such is the the popularity of data linkage that it has spawned a bi-annual conference dedicated to the subject, organised by the International Population Data Linkage Network.

1.4.2 EHR Data Quality

On the assumption that data linkage can make additional data available to fill important gaps in non-health or diffrent spheres of health variables in EHR research, there remains the question of the validity of EHR datasets to deliver data suitable for addressing health specific questions. Several studies have examined the the topic of EHR data quality in secondary use in detail. Botsis and Taxiarchis [78] reported issues they encountered during a survival analysis study for pancreatic cancer. Utilising the EHR data warehouse at the Columbia University Medical Center, they attempted to identify pancreatic cancer patients on the basis of ICD-9 codes. Of 3 068 patients with such a code, an examination of the corresponding pathology reports showed that 1 479 (48%) did not show any documentation consistent with ICD-9 codes for pancreatic malignancies. Of the remaining 1 589 patients, only 522 had sufficient clinical data regarding the progression of the disease to be eligible for their inclusion criteria. Aside from missing and incorrectly coded data, the authors also reported that inconsistencies and inaccuracies such as contradictory information or

poor granularity of diagnosis were also found to be common occurrences. Looking across different institutions, Chan et al [79] conducted a review of 35 studies for EHR data quality indicators across institutions and found that completeness varied a great deal by institution and medical domain, additionally noting that a great deal or heterogeneity in approaches to measuring data completeness existed, and standard methods to assess data quality in EHR research had not yet emerged. A review of literature regarding EHR data quality assessment methods was performed by Weiskoft and Weng [80]. Amongst producing recommendations consistent with Chan et al regarding the need for consistent quality assessment terminology and methodologies, the authors identified five dimensions of data quality that were frequently recognised as sources of concern for EHR research:

Completeness Are all relevant facts represented?

Correctness Are all concepts represented factual?

Concordance Are there contradictions in the dataset, or contradictions when the dataset is linked to another?

Plausibility Do the represented facts fit with our models of reality?

Currency Does the longitudinal representation of the facts create a plausible sequence of events?

Their analysis further emphasised that completeness, correctness and concordance are the principal, non-reducible entities of data quality (plausibility and currency being agents thereof). Notably, completeness was determined as the most commonly assessed entity of data quality in 64% of articles they reviewed, perhaps indicating a general trend amongst EHR researchers that completeness is the most prominent issue [81].

In most cases, such studies acknowledge that completeness issues often concern the availability of coded data compared to free text notes. Coded data might be described as information recorded according to a controlled nomenclature such as ICD-10 or Read Codes [82], including variables such as diagnosis, medication, ethnicity, age and other common attributes. Data structured according to a well understood data model represent ideal conditions for research, as the manipulation and analysis of this data type are far easier than using unstructured data (see chapter 3). Although many EHR systems offer facilities to encode data in such a way, clinical use of these features tends to be erratic, with many clinicians preferring to record patient encounters partially or fully in free text form. A 2016 review [83] by Ford et al identified a range of reasons for negative clinical attitudes to using coded methods of data entry:

- Text is more expressive [84] and captures more nuanced information about the encounter, such as uncertainty, or corrections to previously recorded information [85,86]
- Text serves as a better reminder to the clinician about the encounter [87]
- Finding appropriate codes is onerous and places a burden on the clinician's time [88, 89]
- Coding systems are not complete, and may not represent unusual symptoms and events [87]
- Text is able to express the clinical deduction process, creating a basis of evidence for an appropriate treatment plan [87]

Additional general factors that might be attributed to inconsistent coded data input may include general organisational issues, such as insufficient staff training, poor interface design and a lack of incentive to use an EHR system's features to their fullest capacity. Further, even supplying basic meta-data to appropriately classify free text input has been known to be problematic (for instance, labelling documents as discharge summaries or appointment letters at the point of upload). For instance, Mikkelsen and Aasly identified problems regarding variable accuracy in the capacity of a Norwegian EHR system to support accurate retrieval of clinical text documents, and a predisposition for clinicians not to enter structured meta-data about document types [90]. In summary, clinicians rarely consider potential secondary use cases when interacting with EHR systems, and generally find free text a simpler, more efficient and flexible way to capture and communicate medical information. Thus the free text portion of an EHR often contains the richest source of information about a patient's true status [81, 91].

In order to execute a study using EHR data, a pragmatic solution to tackle the problems of raw EHR data is for researchers to review and recode the data (both structured and unstructured) themselves. Many studies in EHR research formulate a search strategy for putative patients to include in a study, followed by an individual or team of human coders completing a manual review of every relevant patient record to extract the data points for a study. While labourious, this strategy is most likely to produce the maximum data quality according to Weiskoft and Weng's quality criteria. However, this human resource requirement betrays some of the fundamental promises of EHR research in the 'Big Data' era. Such methods impose limits on the amount of data that can realistically be reviewed, and are likely to become unsustainable as the volume of EHR data continues to accumulate. In recognition of the value of free text and poor scalability of manual

coding, many EHR research projects have chosen to invest in natural language processing techniques, discussed in chapter 2, to tackle the shortfall.

1.5 Conclusion

The sustained global effort to exploit the EHR in research reflects substantial progress made over the last 50 years in the implementation of clinical data systems. In parallel, efforts to standardise clinical concepts across geographical boundaries have borne some fruit, and collaborations between healthcare organisations offering data and researchers wishing to make use of it are commonplace. Confidence is high amongst researchers that the full potential has yet to be unlocked, as evidenced by the proportionate investment by funding bodies and the healthy and growing output of EHR based publications observed today. Nevertheless, many challenges are still to be resolved. The swinging balance of trust between patients and care organisations to meet high ethical standards and information governance policies remains an ever present threat to the future of EHR research, and the ongoing inquisition into the substantial data quality issues suggests that the development of methodologies to enhance and refine EHR research are likely to have a widespread scientific impact for years to come.

Chapter 2

Electronic Health Records and Clinical Natural Language Processing

2.1 Overview

Unstructured data are any data that are not organised in a predefined manner, and are therefore inherently more difficult to analyse. Technically, any persistent data item not defined by a well understood data model may be thought of as unstructured data, such as recorded sounds and images, although unstructured data is commonly thought of as synonymous with textual data[1]. In the course of preparing this chapter, I was unable to identify any studies that had attempted to quantify the volume of unstructured data in the world, or the ratio of unstructured to structured data. One frequently quoted figure usually attributed to Merril Lynch is that 80% of the world's data are unstructured; however this claim doesn't appear to have any empirical basis[2]. Further, a 2011 Science paper by Hilbert and Lopez [92] notes that previous studies that attempt to quantify the world's data in all forms struggle to define what constitutes data consumption, and therefore may produce absurd conclusions such as:

"*... computer games and movies represent 99.2% of the total amount of data 'consumed.'* - Hilbert and Lopez, The World's Technological Capacity to Store, Communicate, and Compute Information [92]"

[1] In many ways, labelling human language as 'unstructured data' is counter-intuitive, as the concept of a language might be conceived as a data model, albeit one not particularly suited for machines.

[2] Prominent industy NLP analyst Seth Grimes attempts to investigate the basis of the 80% statistic in this blogpost https://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/

Regardless, it is reasonable to claim that a large proportion of the world's data are unstructured, and textual data account for a substantial fraction.

Natural Language Processing (NLP) is a broad discipline of computer science concerned with the interface between computers and human language in electronic text form. The field has enjoyed a significant and accelerating popularity in both commercial and research sectors over the last 50 years, owing to its many practical applications in the modern information economy. Given the tendency for data generated by clinicians to be in textual form (see chapter 1) and the ease of unstructured data capture, clinical applications of NLP have also emerged as an active research area. Examples of NLP tasks and their applications within the clinical domain include:

Summarisation Document summarisation is the process of converting a document into a smaller document, removing superfluous information in respect to a given objective. For instance, a summarisation methodology of the textual content of a large EHR record may seek to distil the information into an easy to digest summary for different medical specialities

Optical Character Recognition Here, the objective is to process an image of text and extract characters, words and other language constructs into a desired encoding, such that the content is available for downstream tasks. For instance, when one health unit desires to share information with another, images of documents such as health questionnaires are often scanned and uploaded. Since the original text information is not available in a character form, OCR offers a means to attempt the reconstruction of this information

Document Classification It is often useful to be able to infer what categories a document belongs to automatically. For instance, when documents are uploaded into object stores, they may not be correctly 'tagged' by users in order for other viewers to efficiently find them and other processes. A document classification task seeks to assign a given a document to a predefined class (for example, to differentiate between appointment invitation letters and discharge summaries). Note, in this context, a 'document' can refer to multiple language constructs, such as a sentence, a paragraph or an entire manuscript.

Information Retrieval (IR) This task is concerned with identifying relevant documents subject to a user's intent. For example, a user will specify a query relevant to their intention (often composed of a mix of keywords and facets of structured information). IR is the task of calculating the relevancy of every document in a set

with regard to the users query, which can then be used to prioritise the order in which the user reviews the results

Information Extraction (IE) IE is a broad subfield within NLP which seeks to produce a range of structured outputs from text by making use of a wide range of linguistic concepts. This in itself may be composed of additional NLP subtasks. For instance, Named Entity recognition seeks to identify and map entity mentions within a text to predefined concepts, whereas relationship extraction aims to identify the semantic relationships between such entities (such as verbs). By attempting to extract highly granular information, such as the complete set of subject-predicate-object triples from a document, IE approaches can facilitate a broader range of downstream tasks, such as knowledge base population. IE is discussed further below.

Although sporadic references exist to utilising IE over medical records exist as far back as 1970 [93], general interest in the field has predominantly mirrored that of EHR uptake, with a notable rise in biomedical publications referencing NLP since the year 2000 onwards. A systemic review of ninety-seven studies using EHRs to identify patient cohorts by Shivade et al [94] found that forty six had used NLP to supplement data obtained from structured sources. Given the need to convert unstructured data into coded data, a significant focus of clinical NLP concerns two tasks, IR and to a greater extent, IE. IR is principally concerned with efficiently finding documents according to some search criteria, and ranking them according to their relevance. The purpose of IE is to produce facts contained within documents. Accordingly, IE is a substantially harder problem, where solutions tend to be less generalisable than IR methods, since the granularity of the objective is much finer and less amenable to rapidly refining a system to generate new results (often termed the 'serendipity effect' in IR, when one search leads to another). In many cases, IR and IE are used in combination when studying an NLP problem. For example, a researcher may be looking to extract facts about a certain medical concept. However, they are faced with an overwhelmingly large corpus of documents generated from all of an organisations healthcare units, some of which may contain references to their concept. A sensible approach would be to define a broad set of search criteria to efficiently find documents that may include examples of the information they want. By filtering the large corpus with IR techniques, they will be able to produce a much smaller, more manageable corpus to facilitate work on an IE method.

This chapter describes some of the fundamental elements of NLP and how they can be used in basic IR and IE tasks. This chapter also summarises modern approaches to clinical IE problems, and reviews a selection of clinical IE systems that have emerged over

the last ten years. Finally, it closes with an analysis of the trajectory of progress in the field, highlighting opportunities for methodological development.

2.2 Essentials of Clinical Natural Language Processing

A given NLP task may be composed of a number of smaller processing tasks, each contributing to the creation of new 'features' that can be used elsewhere to contribute to the achievement of the overall goal. Modularising the higher order objective in this way facilitates the sharing of results and methodologies between NLP researchers, as well as providing the semantics when unpicking the details of complex NLP tasks. In addition, there have been several attempts to formalise standards for NLP methodologies by the development of NLP frameworks. These include the General Architecture for Text Enginnering (GATE) suite, developed by the University of Sheffield [95], the Unstructured Information Management Architecture (UIMA) originally produced by IBM and now under the management of the Apache foundation [96], and the Python based Natural Language Toolkit [97]. These frameworks tend to organise NLP subtasks into 'pipeline' concepts, whereby each subtask is executed in a logical order, often taking the input of the previous subtask as its input. Although NLP has many sub-disciplines, the scope of this thesis only encompasses IR and IE.

In clinical NLP, it is important to consider Zellig Harris's theories of sublanguages [98, 99]. Amongst other conclusions, these held that a sublanguage is a restrictive set of grammatical components, most often found in technical domains. A degree of assumed knowledge by the reader might allow the author to forego many of the common grammatical rules required to produce grammatical language according to the super-language. Friedman et al [100] provides a detailed examination of Harris's theories applied to the clinical domain, noting features of sublanguages that are readily observable in clinical text. For instance, in a given corpus of normal English, some words are more likely to co-occur than others (e.g. 'patient' and 'doctor' are more likely than 'lamppost' and 'teacup'). This high likelihood of co-occurrence indicates a low information content. In sublanguages, which are predominantly concerned with efficiency of communication rather than grammatical perfection, such features allow authors to ignore many regular grammatical structures. If clinical text was grammatical in the standard English sense, the word 'patient' would co-occur with most other terms in the text (given that most syntactic relationships would have the patient as the subject - 'the patient suffered from X'; 'the patient was taking Y'). This low information content allows authors to omit references to the patient and other entities in many circumstances. For instance, it is common to

observe highly telegraphic sentences in clinical notes, such as “Slept though night.” and “Took medication as directed in morning.”.

By definition, sublanguages should have a more restrictive grammar and subset of legal relationships between entities. In theory, this should make them more amenable to NLP, owing to their lower overall complexity. However, in practice, many irregularities and idiosyncrasies occur. For instance, the variable medium and context in which they are presented to the reader can influence their interpretation, such as a discharge letter with several paragraph headers inviting a description of past and present patient status, or an online free text form field requiring a description of current medication. Below, we describe of the most frequently encountered subtasks in clinical NLP, their application in clinical contexts and the problems posed by the clinical sublanguage.

2.2.1 Tokenisation

Tokenisation is commonly the first process in many English language NLP tasks. Here, a string of text is broken down into its most elemental components (‘tokens’). Generally, this will be individual words, but may also produce tokens from punctuation marks, numbers, negative constructs and other grammatical features, depending on the type of tokeniser used (table 2.1).

I	couldn’t		attend	the	co-ordinator’s				lunch	.
[I]	[could]	[n’t]	[attend]	[the]	[co]	[-]	[ordinator]	[’s]	[lunch]	[.]

Table 2.1: Example of tokens from a text string

Many implementations of tokenisers exist. Although conceptually simple, selection of an appropriate tokeniser is highly dependent on the required task. For instance, Akkasi et al [101] found difficulties in the successful tokenisation of chemical entities when using common tokeniser implementations, and developed a new system accordingly. Similarly, the Textrous! system [102] uses a custom tokeniser for optimal tokenisation of gene names and symbols. Tokenisation is also an important component of many IR tasks. For example, when a document is ’indexed’ (i.e. uploaded) into a Lucene based search engine, it is tokenised via a preconfigured method in order to populate an inverted index. The method of tokenisation defines many aspects of how a document can be searched.

2.2.2 Sentence Boundary Detection/Splitting

The Merriam-Webster [103] definition of a sentence is

“*. . . a word, clause, or phrase or a group of clauses or phrases forming a syntactic*

unit which expresses an assertion, a question, a command, a wish, an exclamation, or the performance of an action, that in writing usually begins with a capital letter and concludes with appropriate end punctuation, and that in speaking is distinguished by characteristic patterns of stress, pitch, and pauses."

As its name suggests, Sentence Boundary Detection refers to the task of accurately partitioning a document into distinct, non-overlapping sentence boundaries. As with tokenisation, defining sentence structures provides useful meta-data information for other downstream processes. For example, one might wish to look for grammatical relations to a specific token, but only within the sentence that the token appears (or the surrounding sentences). In NLP complexity terms, this task is concerned with identifying syntactic structures that appropriately delimit sentence units. In perfect grammatical English, this may involve differentiating between the use of capitalisation and the full stop punctuation mark to define sentence boundaries and other valid uses, such as in proper nouns and abbreviations. However, sublanguages often do not conform to strict grammatical rules, meaning sentence splitting is often a non-trivial activity. For instance, social media usage tends to follow seemingly protean rules of grammar, prompting researchers to create domain-specific sentence splitters [104]. Similarly, clinical language often makes extensive use of lists and abbreviations in various states of capitalisation. A review by Griffis et al [105] identified little research in the domain of clinical Sentence Boundary Detection, and that many of the commonly used algorithms that are designed for well formed English perform badly on clinical text.

2.2.3 Dictionary Lookup, String Matching and Lexical Normalisation

The presence of certain keywords and phrases in a given text string are often valuable pieces of information when considering the overall objective. Therefore, simply tagging such information is a common way of making the presence of relevant entities available to other resources. For instance, if a task seeks to extract medication information about an individual, the term 'paracetamol' in a sentence may flag it as a candidate for containing further information, such as a predicate linking the medication to a dosage value. One of the most common sources of dictionary lookup is the Unified Medical Language System Metathesaurus (UMLS), a project aimed at integrating and mapping concepts between key medical nomenclatures. Supported at the National Library of Medicine (NLM), the 2016AA release of the UMLS Metathesaurus contains 9 080 363 English language terms. Given its sheer size, many projects only make use of a subset, in order to address computational memory limitations. Examples of systems that make use of UMLS are described

in 2.3.8.

Although they are in English, UMLS terms aren't intended to represent part of the natural spoken or written language, which creates substantial problems in matching the natural language found in clinical text to UMLS entries. This in part might be addressed via configuring simple lookup systems for an exact or approximate string match of the input string. More complex approaches to dictionary lookups attempt to resolve a variety of forms that a given entity can take in natural language. For instance, "Sickle Cell Anaemia" might also be written "Anaemia, Sickle Cell". Appropriate mapping to a dictionary term requires a process called lexical normalisation, which remains an active area of clinical NLP research. Leaman et al [106] note that progress in clinical text lexical normalisation is substantially lacking compared to other domains, due to the additional linguistic complexities posed by the sublanguage. Systems that attempt to use lexical normalisation are further discussed in 2.3.8.

While not the focus of this thesis, the concept of approximate string matching is worth examining in a little more detail, due to the significance of the technique in the fields of text de-identification, record linkage, molecular sequence alignment and spell checking to name but a few. The problem is often conceptualised as 'edit distance', in that the similarity of two strings can be parameterised according to some function. Common approaches in NLP include Levenstein distance [107] (later expanded to Damerau–Levenshtein distance [108]) and Jaro–Winkler distance [109] amongst others. In practical applications, approximate string matches will provide better results when documents contain misspellings or have dialect variations (e.g. American English 'color', vs British English 'colour'). Conversely, partial matching can yield inaccuracies if the similarities between two strings meet the match threshold, yet are genuinely different concepts (e.g. the drugs 'Clozapine' and 'Clonazepam').

2.2.4 Co-Reference Resolution

In linguistics, a referent is the subject of an expression. Co-references occur when multiple expressions have the same subject. The NLP task of co-reference resolution, therefore, is the accurate assignment of multiple expressions to the correct subject. For instance, the following snippet:

"*... the patient suffers frequent headaches. He often takes paracetamol for this.*"

contains two expressions. The subject of both expressions, 'patient' and 'He' refer to the same individual. Therefore, an NLP algorithm that can resolve these co-references

would also be able to deduce that the patient is taking paracetamol for heachaches (rather than another malady that might be referenced elsewhere in the document). Co-reference resolution is known to be a hard problem in general NLP research, although has received little attention in clinical NLP [110]. This may be in part due to the telegraphic nature of the clinical sublanguage, in that most assertions are likely to concern the patient to which the document is attributed. However, situations do occur where such an assumption is invalidated, such as the discussion of care team staff, or of family members in the context of disorders with a known genetic basis.

2.2.5 Morphological Segmentation/Stemming and Lemmatisation

Often, concepts are expressed in a variety of inflections, such as past or present tense, or in plural form. In order for an NLP system to exhibit a degree of robustness under different inflections, it must be capable to handling such situations. These tasks entail the mapping of the various inflections of words back to their elemental constructs. Lemmas and stems refer to slightly different concepts. A stem of a word is always an initial substring, computed using a heuristic to remove appropriate parts of the original form. The most commonly implemented stemmer uses Porter's algorithm [111]. An example is given in table 2.2.

The	patients	haven't	been	seen	by	the	nurses	yet	.
[The]	[patient]	[have]	[been]	[seen]	[by]	[the]	[nurs]	[yet]	[.]

Table 2.2: Example of stemming with Porter's algorithm

Such an example shows a range of problems with the stemming concept. The negated form of 'have' is lost, changing the meaning of the sentence, and the word 'nursing' is reduced to 'nurs', which has no meaning, therefore cannot be identified as an entity without additional work.

A more sophisticated approach is to specifically attempt to identify the base form of a given word, or lemma. Implementations of lemmatisation algorithms include the NLTK WordNetLemmatizer [97], and the Stanford CoreNLP lemmatiser. Based substantially on the work of Minnen et al [112], the Stanford Lemmatiser uses an expansive series of grammatical rules to determine the correct lemma. Table 2.3 shows the same statement processed via the Stanford CoreNLP package, giving a more satisfactory result:

Nevertheless, stemming remains a popular approach in IR tasks, where the simplicity of the concept confers advantages when speed and scalability are important factors. For instance, stemming is the prevailing method of handling word morphology in the popular

The	patients	haven't		been	seen	by	the	nurses	yet	.
[the]	[patient]	[have]	[not]	[be]	[see]	[by]	[the]	[nurse]	[yet]	[.]

Table 2.3: Example of lemmatisation with the Stanford CoreNLP package

Lucene search engine and its various implementations. Lemmatisation research specific to the clinical English sublanguage appears to be limited.

2.2.6 Part-of-speech Tagging

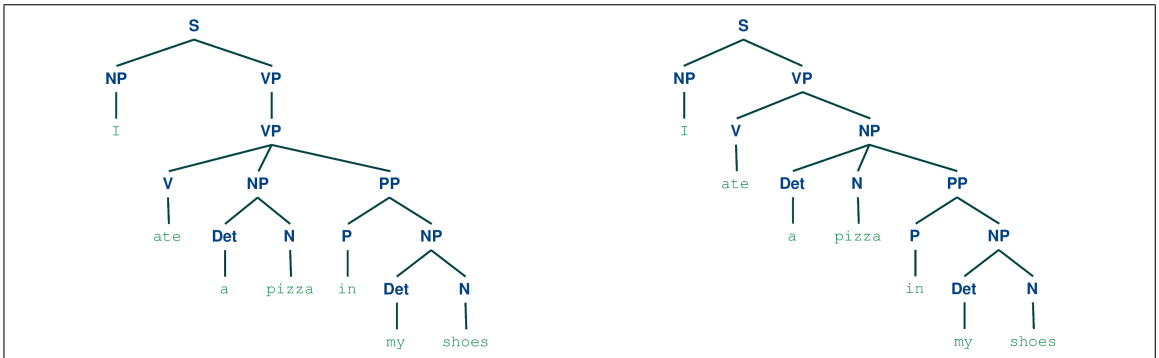
Parts-of-speech (POS) refer to word classes, such as noun, verb, adjective, number etc. The role of a POS engine is to correctly assign POS tags. As with many other NLP subtasks, POS tags can add an additional layer of information for use in downstream processes. For instance, consider the sentences given in table 2.4, annotated according to the Penn Treebank [113] model of POS tags.

David	is	cooking	.
Victoria	is	cooking	.
[NNP]	[VBZ]	[VBG]	[.]
[proper noun]	[verb - 3rd person]	[verb - gerund]	[.]
[singular]	[singular present]	[or present participle]	[.]

Table 2.4: Part of Speech tagging of two simple expressions

Both sentences induce the same pattern of POS tags, identifying both David and Victoria as proper nouns. A simple NLP system might therefore utilise such information to identify other proper nouns that follow the same pattern.

As with many NLP tasks, much of the research effort for POS tagging has been developed on well formed grammatical structures. Modern efforts show state-of-the-art classification performance with the use of recurrent neural networks [114]. However, such methods have not yet filtered through to the clinical NLP domain, and older approaches demonstate somewhat poorer classification performance when applied to clinical text [115]. While the main public effort to develop a clinical POS tagger comes from the cTAKES architecture [116], Fan et al noted that a cross institution evaluation of the cTAKES tagger performed less well than the generic Apache OpenNLP tagger [117]. Both taggers showed a marked classification performance loss between institutions, suggesting generalisability issues in clinical POS tagging. However, domain adaption shows promise for future directions of clinical POS tagging (see section 2.3.5.)



(a) Sensical interpretation

(b) Nonsensical interpretation

Figure 2.1: Examples of sensical and nonsensical interpretations of the same sentence. While trivial for a human to resolve, correct interpretation by a machine requires sophisticated statistical methodologies

2.2.7 Chunking and Parsing

The purpose of chunking is to build a formal structure around a sentence, by dividing it up into high level non-overlapping segments. Typically, this involves the identification and grouping of tokens into noun phrases (commonly abbreviated 'NP'). Chunking algorithms tend to make use of POS tags to identify appropriate text chunks, and use fast, rule based heuristics or transformation-based learning methods as proposed by Ramshaw and Marcus [118]. For instance, table 2.5 shows NP chunks generated by the GATE implementation of the Ramshaw and Marcus base noun phrase chunker.

I	ate	a	pizza	wearing	my	shoes	.
[PRP]	[VBD]	[DT]	[NN]	[VBG]	[PRP$]	[NNS]	[.]
[NP]	[]	⊢ NP		[]	⊢ NP		[]

Table 2.5: Examples of noun phrase chunking

As with POS tagging, chunking can add valuable higher order information to assist with more complex NLP tasks. Savkov et al [119] studied the classification performance of a variety of general English and clinical text specific chunkers over clinical text, showing a wide range of classification performance.

In addition to chunking, parsing (or syntactic analysis) is frequently employed in IE. The objective of parsing is to obtain a dependency structure (or parse tree) by identifying grammatical relationships between lower syntactic units. Parsing is a highly complex task, as many sentences can generate a wide selection of dependencies, many of which are nonsensical to a human observer who would commonly have a great deal of contextual knowledge about the referenced entities (figure 2.1).

Current state of the art parsing research makes use of non-recurrent feed forward neural networks (revisited in chapter 7), such as the 'Parsey McParseface' parser developed by Google Labs [120]. However, validation studies of parsing methods upon clinical text are sparse. As with many NLP subtasks, Jiang et al [121] demonstrated that parsers trained upon general English corpora generalise poorly when applied to clinical text. The authors note that in many cases, the telegraphic nature of clinical text render it unsuitable for general parsing concepts. Nevertheless, a recent review by Velupillai et al [122], spoke in favour of a greater research effort in semantic analysis of clinical text, noting that the most performant systems in the Analysis of Clinical Text shared task in SemEval 2015 made use of such approaches [123] (although also acknowledging that semantic analysis was probably over-engineering for a variety of practical clinical NLP use cases).

2.2.8 Word Sense Disambiguation

Word Sense Disambiguation is the task of accurately identifying the correct sense of a word for a context. For example, 'bark' may be foud in the context of "the dog's bark" or "the tree bark". Word Sense Disambiguation also applies to acroymns. In medical terminology, the importance of Word Sense Disambiguation is reflected by acronym overloading range as high as 81% in the UMLS [124]. A further issue is that many acronyms are invented in an ad hoc manner by care teams to ease internal communication. State of the art methods include the CARD system [125], which utilises machine learning in combination with UMLS to detect widely used and custom abbreviations in a given corpus.

2.2.9 Temporality

Temporal data extracted alongside the event they relate to are highly valuable in order to build knowledge for a range of problems, such as understanding disease progression and response to treatment. Temporal event extraction is the process of identifying temporal modifiers within text, and accurately capturing how they change the human interpretation of text. For example, correctly extracting and assigning the semantic relationships to the temporal modifiers in the text "the patient's MMSE was 15/30 last year. Today it is 13/30." is essential to build a representation of MMSE decline. This is a well-recognised challenge within both general NLP and clinical NLP specifically, to the extent that it has been the subject of an NIH-funded international research competition [126]. Although this thesis does not specifically investigate temporality, we refer to this concept throughout as an important aspect of clinical NLP work.

2.2.10 Negation Detection

Negation detection is the identification of lingustic constructs that reverse the polarity of a statement. In 2002, a seminal paper by Chapman et al [127] identified that a very large proportion of assertions in clinical text contained negation modifiers, and proposed NegEx: a simple regular expression (see 2.3.3) based system for identifying negated entities. Identifying this unusual feature of the clinical sublanguage would prove influential. A review of negation detection and other common clinical modifiers by Meystre et al [128] explored the generalisability of the NegEx algorithm, and found it ranging from 77% precision and 83% recall to 94.3% precision and 94.5% upon different datasets, suggesting negation is not a 'solved' problem. Despite (or perhaps due to) its simplicity and the appearence of other negation solutions, the popularity of the system has endured, as borne out by its integration into a number of NLP systems such as cTAKES [116] and BioLark [129]. In 2009, the algorithm was expanded to include temporality and experiencer (subject) concepts, although classification performance in these new concepts was not comparable with the original NegEx algorithm [130]. Nevertheless, negation detection remains a critical component of clinical IE.

2.3 Approaches to Clinical Information Extraction in Clinical Research

Combining the various subtasks of NLP is a common way to build layers of additional data about the textual component of a document to facilitate IE tasks. For instance, a common task in IE is Named Entity Recognition (NER). NER is the task of identifying real world entities in a document. It is often confused with IE, as NER can also be an end in itself depending on the requirements of an IE problem. However, in clinical IE, NER is often combined with other tasks, such as negation, temporal and subject detection in order to provide additional contextual information for a given named entity [131]. These additional data are often described as 'features', which can then be used in turn for rule-based, ML or hybrid (a combination of the two) approaches for IE. Regardless of the approach used, the availability of corpora of appropriately annotated data for a given NLP task, is usually imperative to support evaluation. This section introduces these two broad schools of thought regarding IE, the statistical evaluation approaches, and the challenges of obtaining corpora for evaluation tasks.

2.3.1 Creating Corpora

Annotated documents are produced via one or more human annotators, who follow annotation guidelines to ensure consistency in the task. Annotations are generated by the use of software specifically designed for the task, such as the BRAT [132] tool, Knowtator [133] or the aforementioned GATE package (Figure 2.2). In essence, annotations are meta-data containers for text to describe any potential feature of it, such as the boundaries of a sentence or a named entity. The purpose of such metadata is to either:

1. represent the intended final state of an NLP process, thereby enabling a quantitative assessment of classification performance

2. in the case of Machine Learning (see below), to supply training data to an algorithm.

In general NLP research, much effort is expended on the production of suitable corpora and their associated annotation guidelines, to provide international standards by which researchers can validate their efforts. The annotation of corpora is generally task specific; for instance, one of the earliest annotated corpora, the Brown corpus [134], has over one million tokens annotated predominantly with part of speech tags from a wide range of American English literature. Later examples include the PENN TREEBANK, offering around 3m words annotated with both POS tags and syntactic structure [135, 136]. As interest in general computational linguistics and NLP has grown, a wide range of other resources have emerged. Specifically regarding IE, an important early corpora to emerge originated from the series of seven Message Understanding Conferences. Originally devised to promote the automated interpretation of military communication, the legacy of the Message Understanding Conferences helped to standardise activities such as evaluation metrics and the semantics regarding our current interpretation of co reference detection and NER. Corpora specific to clinical NLP are discussed further in section 2.3.7.

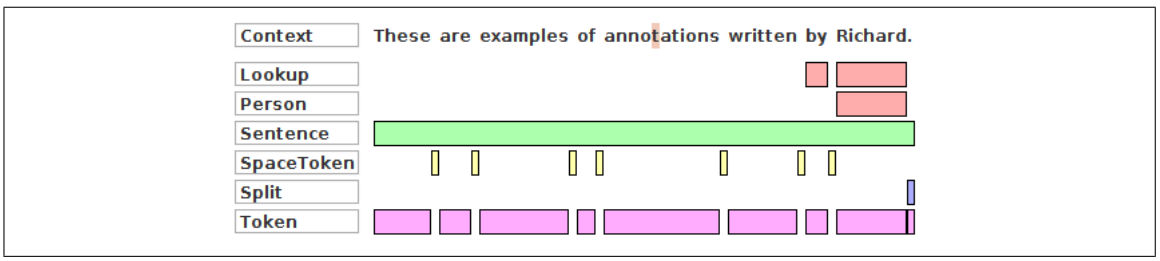


Figure 2.2: Examples of annotations in GATE

2.3.2 Evaluation Metrics

Using an annotated corpus, it is possible to conduct classification performance evaluations for a given NLP subtask or IE process. Statistics in IE and IR tasks are generally reported

in terms of classification performance of the model in a set of test data. For a given corpus, let TP = the true positive count of a certain entity, TN = the true negative count, FP = the false positive count. Precision (or positive predictive value) is therefore defined as:

$$\frac{TP}{TP + TN}$$

Recall (or sensitivity) is defined as:

$$\frac{TP}{TP + FN}$$

Together, these two metrics offer allow an observer to interpret the two most important classification performance aspects of a given system. Often, the harmonic mean of these two values is also reported (the F1 measure), allowing for a meaningful comparison between methods that may perform better and worse with regard to the individual precision and recall metrics. Error bars may also be included using a binomial proportion confidence interval to provide an indication of sampling error.

Regarding the development of data annotated to form a gold standard (for instance, by using aforementioned guidelines developed for a specific task), it is often helpful to understand how consistently such guidelines have been applied. This is often performed by having two or more humans annotate the same information independently, and calculating inter-annotator agreement (IAA) statistics. The simplest statistic is merely the percentage agreement between two annotators. However, a more informative statistic is Cohen's Kappa, which also takes into account any agreement that might occur by chance. Cohen's Kappa is defined as

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o is the observed agreement and P_e is the probability of chance agreement. The result is normalised between 0 and 1. An outcome close to 1 indicates a high degree of 'true' agreement, whereas a result close to 0 indicates most agreement is by chance alone.

IAA analysis offers an insight into how well understood a concept is within a given domain. A low IAA result indicates disagreements about conceptual elements of an entity, and that further debate/guideline development or annotator training is required.

The resulting corpus, composed of annotations produced by multiple annotators can then be used to define a 'gold standard'. Here, the instances that annotators agree can be used to indicate the upper bound of what the classification performance of an IE approach can achieve - by definition, an automated approach to a problem can only perform worse, or as well as, the standard by which it is assessed.

2.3.3 Rules based

Rule based approaches (also referred to as 'knowledge engineering') use purpose built computational data structures and manipulations to define patterns of interest, producing a deterministic output. At their most elemental, regular expressions (RegExs) are created to act over character strings, triggering when certain criteria are met. RegExs have proven to be extremely useful in a host of string manipulation contexts, such that some form of RegExs processing is available in most of the major programming languages. A Java example is given in listing 2.1.

Listing 2.1: A simple function utilising a RegEx in pure Java, replacing all numeric digits with the letter "X"

```
public static String maskNumber (String input){
   return input.replaceAll("[0-9]","X");
}
```

RegExs are limited in their capability to work over complex grammatical structures, such as those generated by NLP subtasks. To this end, a variety of task specific programming languages have arisen, such as Java Annotations Pattern Engine (JAPE) [137], built in to the GATE framework [95], and the RUTA [138] engine, which is part of the UIMA framework [96]. Such languages allow language engineers and informaticians to efficiently manipulate the features produced by other NLP components. A simple JAPE example is given in listing 2.2.

Listing 2.2: A simple negation rule written in JAPE. The desired annotation pattern is written first, and the output is defined after the "–>" symbol

```
Phase: First
Input: Lookup Token
Options: control = appelt

Rule: CancerNegation
(
  {Token.string == "not"}
  {Lookup.majorType == "cancers"}

)
:cancerNegation
-->
:cancerNegation.Cancer = {negation = "true"}
```

Rule based approaches make use of the aforementioned NLP subtasks, such as extensive use of dictionaries and knowledge of appropriate parse trees to find general lingustic patterns regarding a specific task. Hand crafted rule based approaches are often effective for many problems, and owing to their simplicity, have the advantage of a transparent logical interpretation to human observers. This enables researchers to gain an intuitive understanding of why errors occur. Rule-base systems have a time honoured effectiveness in many complex NLP tasks. For instance, the rule based HeidelTime system came first in TempEval2, a shared task designed to test temporal relation detection [139]. Nevertheless, many NLP frameworks require a degree of technical knowledge and experience to employ rules successfully. In addition, if a task requires large numbers of rules, they can become unwieldy to manage.

To counter some of these limitations, some investigators have sought to automatically determine appropriate rules for a given problem. This a task known as supervised rule induction, a form of ML. Here, an algorithm is employed to induce rules, based upon annotations made in a corpus of training data. This process has been employed successfully to extract ICD-9 codes from radiology reports [140], using the RIPPER algorithm [141]. RIPPER in turn makes use of the more general C4.5 algorithm to build readily interpretable decision tree classifiers. Such methods are a useful bridge between some of the more opaque ML methods described below and hand crafted rules, although often fail to

provide the classification performance of more advanced ML methods in NLP tasks.

2.3.4 Machine Learning

In academic circles, ML approaches are an increasingly popular way to approach many types of NLP task. Nadkarni et al [142] attribute this to the limitations of hand crafted rules described above (although researchers from IBM also suggest that this is at least partly attributable to academic biases regarding the low research potential of rules [143]).

Regarding IE, the most popular class of ML algorithms are known as discriminative models. In these, training data takes the form of annotated corpora in NLP. One of the most frequently employed examples of such models over the last ten years are called Support Vector Machines (SVMs), which are used extensively throughout this thesis. SVMs aim to assign one of two classes for a given instance (i.e. samples). To illustrate their function, I use data from a classic experiment - Fisher's Iris data [144]. Here, the objective is to discriminate between two species of Iris plants, I. setosa and I. versicolor (the original experiment considered three species, but for simplicity I shall only consider two at first). The first step in building an SVM classifier requires the acquisition of labelled training data. In total, there are 50 samples from each of the two classes. For each instance, four attributes are recorded - sepal length, sepal width, petal length and petal width. At first, we shall only consider two - sepal length and petal length. If we produce a scatter plot of the data, we obtain the graph given in Figure 2.3.

Here, the two attributes produce clearly distinguishable clusters for each of the two species. The distance between the two clusters is called a margin. Many potential margins exist (Figure 2.4), although one will be maximal, in that it produces the greatest Euclidean distance between the two clusters. This is created by identifying the support vectors - the instances on the edge of each cluster that are nearest to the opposing cluster.

The intuition behind SVMs is that the line that is the midpoint on this maximum margin differentiates the two classes the best. Therefore, if we were to predict the class of a new, unlabelled sample, an SVM classifier would choose a class depending on where the new sample appeared in relation to this line.

In more complex datasets, we might have more than two attributes. For instance, if we were to include sepal width as a third attribute, we would need to plot the same graph with a third axis. At this point, we need a two dimensional line, known as a plane. Such a concept can be extended to n attributes, at which point, we refer to the plane as a hyperplane.

In the real world, data often cannot be separated by a linear hyperplane. For instance, consider the same plot, but using the much more similar species of I. versicolor and I.

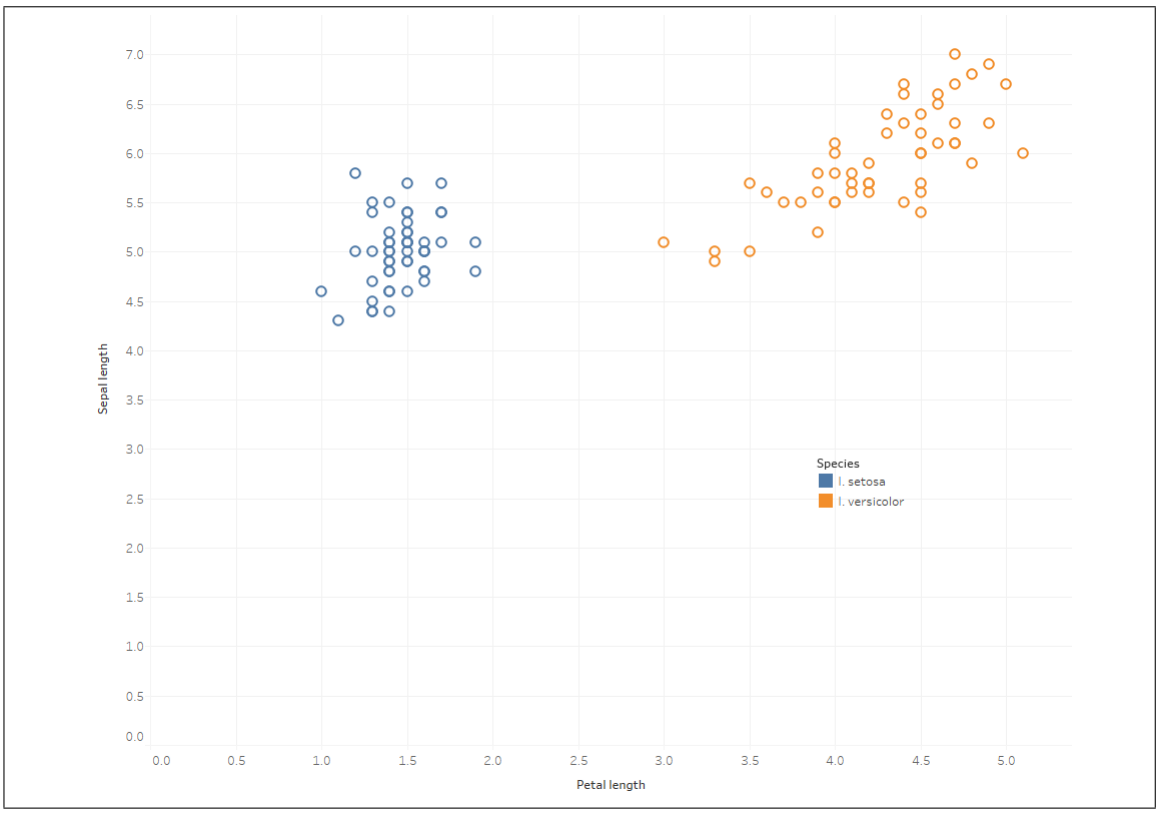


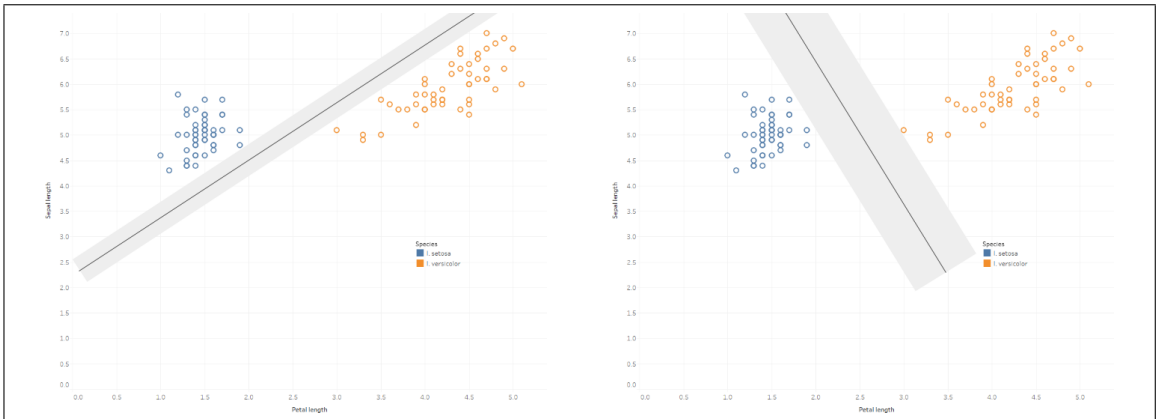
Figure 2.3: Petal length and sepal length of two species of Iris, forming readily distinguishable clusters. Data from [144]

virginia (figure 2.5). The power of SVMs come from using kernel functions, to convert the simple n dimensional space to a much higher dimensional *feature space*, where a linear hyperplane may be found more easily. Commonly used kernel functions in NLP applications are the polynomial and radial basis function.

Even with kernel functions, perfect separation of labelled classes is not always possible. Without further parametrisation of the SVM, the resulting model may be overfit to the training data, meaning that it generalises poorly when attempting to classify unseen instances. To this end, a cost parameter C is introduced. C allows some degree of misclassification by permitting soft margins, in pursuit of a more generalisable overall model.

Although SVMs are binary classifiers, in that they can only discriminate between two classes, several approaches have emerged to generalise the algorithm such that multi-class classification is possible. The most computationally efficient is "one vs all", wherein one class is tested against all other classes combined together. For a given unseen instance, the class that offers the greatest distance from the hyperplane is selected. Similarly, "one vs one" tests each class against every other class in turn. While this may result in better overall classification performance, it is more computationally expensive.

SVMs have been widely used in IE applications, where attributes are derived from



(a) Small Margin (b) Large Margin

Figure 2.4: Margins separating two species in Fisher's Iris dataset. The grey areas represent two margins, one large, the other small. The black lines represent two theoretical hyperplanes. Theoretically, the larger hyperplane should be a better classifier of unseen data.

the NLP subtasks as described above. This commonly includes simple attributes such as the 'bag-of-words' approach, where word tokens are included in the model without any attributes describing the syntactic relationship between them. However, applying further semantic analysis in pre-processing may increase the classification performance of SVMs and the range of tasks for which they are suitable [122]. Their popularity might be attributed to good classification performance even when training data are sparse (as is often the case due to lack of annotated corpora). Although classification performance may be improved via providing additional training data, some have theorised that the efficiency of training may be increased by the use of active learning style approaches [145]. Here, the theory is that instances that lie close to the hyperplane of a given model contain the richest information, as the model has the most difficulty differentiating to which class they should belong. Therefore, any activities to increase training data for the model should prioritise the annotation of such instances.

Another ML approach well suited to NLP problems is Conditional Random Fields (CRFs). CRFs are a sequence based classifier which make use of logistic regression to probabilistically classify a given input, based upon a sequence of previous inputs (for instance, a sequence of words). The consideration of non-independent attributes make them a natural choice for NLP. For instance, in NER style tasks, CRFs have been shown to outperform SVMs in clinical contexts [146].

Although ML approaches hold great promise in terms of circumventing the issues with rule based systems, they also introduce complications of their own. Aside from requiring a different set of skills from a language engineer to obtain the best results, they are generally

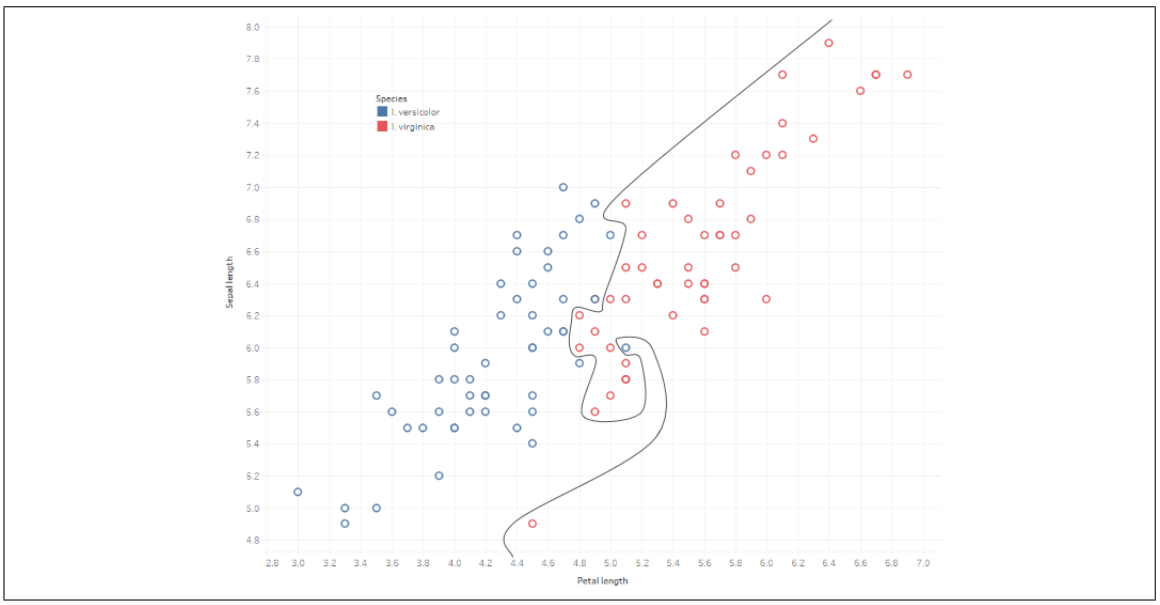


Figure 2.5: Two species in Fisher's Iris dataset, separated by a non-linear margin

limited by the availability and ease of creating appropriate training data [147,148]. Many NLP systems employ both rules and ML, according to the complexity of the task and the availability of resources.

2.3.5 Domain Adaption

Many state-of-the-art methodologies in NLP are predicated on ML approaches, which are dependent on the availability of domain specific annotated corpora to fuel the development of high accuracy statistical models. Generally, such corpora originate from standard English sources, such as the Brown Corpus (see section 2.3). When such a model is applied to out-of-domain corpora, such as the clinical sublanguage, significant drops in classification performance are usually observed. For IE subtasks, such as POS tagging, such errors can propagate upwards to the overall objective, adversely affecting the contribution of such subtasks to obtaining higher semantic layers of understanding. At its simplest, domain adaption involves retraining statistical models with additional data annotated from the new domain to provide better overall classification performance. However, since the availability of such corpora are extremely limited for the clinical sublanguage, researchers have devised more sophisticated methodologies to address this issue. Ferraro et al [149] estimated accuracy drops in the region of 8.5-15% of general purpose POS taggers when applied to clinical text. To counter this, the team developed a technique called ClinAdapt. Here, a new set of training data is constructed in a three step process. First, a set of clinical language sentences are processed with a standard POS tagger (in this case, one from from OpenNLP package trained on Wall Street Journal text). Secondly, a set of clinical terms and their associated POS tags are obtained from the UMLS SPECIALIST termi-

nology (terms were selected on the basis of having unambiguous POS tags assigned to them). Finally, a transformation-based learning algorithm is used to derive the rule from a pre-existing set of rule templates that corrects the most POS tags over a training iteration. Multiple iterations of training are used, as a selected rule from a previous iteration may introduce new errors elsewhere in the training corpus, ending with the selection of a rule that produces the fewest net errors over the entire corpus. A cross validation assessment on two clinical corpora of the ClinAdapt method showed an accuracy of 93.2–93.9%, compared to the best general purpose tagger with 88.6%.

2.3.6 EHR Implementation Considerations in IE

Semi-structured data might be described as having properties of both unstructured and structured data. Here, data are predominantly unstructured, but have useful meta-data available. By this definition, a rarely acknowledged facet of unstructured data in EHR systems is that they might be considered 'semi-structured', as many structured types can be linked to them via relational database queries. For instance, one might reasonably expect to be able to link uniquely identifying information such as a hospital ID number, demographic information such as gender, age and ethnicity, and specific document labels such as 'Ward Note' or 'Discharge Summary' to any given document in an EHR system. Such information confers a range of additional resources that might be brought to bear in a range of clinical NLP tasks, such as NER and temporality. For instance, timestamps are usually automatically recorded whenever a piece of clinical text is generated or a document is uploaded. Nevertheless, to simplify the semantics in this thesis, we use the terms 'unstructured data' to refer to EHR text in isolation, and 'semi-structured data' to refer to EHR text enhanced with metadata, when appropriate.

Employing IE within EHR datasets tends to be a trade-off between precision and recall. Depending on the requirements of a task, this can facilitate looking at clinical documents as a longitudinal patient narrative. This has implications for designing an IE system. For instance, in cohort identification, merely identifying that a patient has received a certain diagnosis may be sufficient. Depending on the frequency of the diagnosis of interest, it is possible to tune an IR or IE approach to favour either precision or recall. As there may be multiple independent references to a given concept across a single patient's record, there can be multiple opportunities for an algorithm to capture a piece of information. Such a situation would favour a high precision approach. Similarly, an erroneous data point generated by poor precision may only need to occur once for it to enter a final dataset. In addition, classification performance is not likely to be uniform across all entities within a class, as not all entities within a class are written about in the same manner. For example,

clozapine is a heavily monitored drug with potentially fatal side effects, thus requiring patients to visit a hospital many times over the course of the treatment. Clinicians may write very differently about relatively safe drugs, or even not include references to drugs not requiring a prescription at all. In order to work around these issues, study-specific post-processing procedures and testing are often needed to maximise data quality. Generally this will involve an iterative process of filtering annotations and aggregation of extracted data points within suitable time windows, and further manual annotation in the context of the research question to ensure the data is fit for purpose.

2.3.7 Shared Tasks in Clinical Natural Language Processing

The production and sharing of clinical corpora is problematic for a number of issues. Primarily, patient privacy concerns and fears of inadvertently exposing inferior, clinical practice potentially subject to litigation have lead to a disinclination of institutions to expose data to external observers [150]. This is further compounded by the difficulties in producing high quality clinical corpora, due to the high cost of training of clinical domain experts, and subsequent time investment required on their behalf [151,152]. Finally, where annotated resources do exist, Chapman et al [150] note that a lack of conventions between clinical NLP research groups means that such corpora are often annotated according to different guidelines, affecting the ease of validating tools in different contexts. Nevertheless, several initiatives have arisen to offer annotated clinical corpora to NLP researchers, in order to establish state of the art practice in several tasks.

Informatics for Integrating Biology and the Bedside

The Informatics for Integrating Biology and the Bedside (I2B2) organisation was a National Institute for Health funded body with a remit to progress translational medicine in the United States. Amongst its activities were the organisation of an annual or bi-annual set of shared tasks in clinical NLP. These tasks were based around annotated clinical corpora that national and international research groups could apply to access, for the purpose of developing and testing NLP applications for certain goals. Beginning in 2006, such tasks have included clinical text de-identification (2006, 2014), identification of obesity and its co-morbidities (2008), extraction of medication information (2009), relationship extraction between medical entities (2010), co-reference resolution(2011), temporal relationship extraction (2012) and diabetes/heart failure risk factors (2014). Results of each event were shared in special issues of the main medical informatics journals.

The third I2B2 Shared Task ran from June to August 2009, with the results published in 2010 [153]. Here, the challenge was to extract medication entities and associated modi-

fiers from 1 243 discharge summaries, of which 547 were released to twenty teams around the globe prior to the evaluation period. Important modifiers included temporal factors such as duration and indication. Of the 547 summaries, 251 were collectively annotated according to guidelines specified by the organisers. The corpus itself contained a mixture of 'running narrative' style text, and 'list structure' style text, each containing medication entities designated for extraction.

The top performing attempt utilised hybrid approaches, making use of SVM and CRF based classifiers [154] and specifically constructed rules to achieve an F-measure of 0.857, although the authors did not provide details on the relative contribution of the ML and rule based components. The second most performant system made use of an existing rule based piece of medication extraction software called MedEx [155], and expanded it with some additional rule based components to meet the requirements of the annotation guidelines. This system performed marginally worse with an F-measure of 0.821. In third place, another rule based approach scored an F-measure of 0.812 [156], with the authors noting the rapid speed at which they were able to develop rule based systems, compared to the relative problems of obtaining sufficient annotated data to enable a ML approach.

Of the top ten submitted systems, all performed worse on running narrative text compared to list style by a large F-measure margin, in the order of 0.2. Error analysis surmised that this was mainly due to the existence of the aforementioned modifiers in natural language, requiring substantially more complex syntactic processing. The top performing system delivered an F-measure of only 0.656, compared to 0.873 on the list style documents. Such a result suggests it may have been somewhat inappropriate to consider the two tasks as a single problem (apparently the distribution of documents classed as each type was skewed in favour of the list style). Presumably in doing so, the desired goal of the task was to encourage the development of generalisable NLP systems capable of handling multiple document types. Nevertheless, such a result reflects the previously identified telegraphic and diverse nature of the clinical sublanguage.

The 2014 I2B2 event concerned a retrospective of recent years of clinical NLP research, and tasked participants with a practical application in identifying risk factors for Coronary Artery Disease (CAD) over longitudinal patient narratives [157]. The corpus was composed of 1 304 patient narratives from 296 patients, with 60% of the corpus made available to twenty teams, who submitted a total of 49 approaches. A small portion of the corpus was annotated according to CAD risk factors, along with the risk factor severity, and made available to the teams.

Of the 49 systems submitted, the best classification performance was achieved by a system developed by the NLM [158]. The system was remarkable for its simplicity, attain-

ing an F-measure of 0.9276. To increase the amount of training data, the NLM team chose to annotate for themselves the unannotated portion of the corpus that was made available to them. This was supplemented with an external dictionary of risk factors and some rules appropriate to the task (including the ConText algorithm for negation detection). Finally, the team built a series of simple SVM classifiers to meet the requirements of the task. Although other submissions admittedly came very close to the top result, many of the systems featured a range of complex components [159] or ensemble (multi-algorithm) ML methods [160]. Overall, these added little to the overall classification task.

Centers of Excellence in Genomic Science Neuropsychiatric Genome-Scale and RDOC Individualized Domains

Centers of Excellence in Genomic Science (CEGS) is a program sponsored by the National Institute of Health to facilitate interdisciplinary exploitation of genomic research. Recognising the value of clinical data in developing phenotypic profiles of patients, CEGS organised three shared tasks for the clinical NLP community to run throughout 2016, with corpora made available for each. The tasks were:

- Clinical text de-identification over approximately 1 000 psychiatric records [161]. This task was further slip into two streams, asking researchers to offer solutions 'out-of-the-box' (i.e. where the accuracy of systems were evaluated without access to any prior training data), and where researchers were given an opportunity to train models on a subset of the annotated documents, prior to evaluation on a test set. The results of the evaluation suggested that the proposed de-identification systems did not generalise well without the benefit of in-domain training data, with the best system producing an F1 score of 0.799. Where training data was available in the second stream, the highest scoring system produced an F1 of 0.914 [162], suggesting clinical text de-identification is still a difficult problem.

- Symptom severity classification based upon the for the Research Domain Criteria framework (RDoc) over 816 documents.The RDoc is a non-diagnostic set of guidelines for evaluating mental health in respect to dysfunctional psychological and biological systems. This task was framed as a document classification problem, with data sourced from initial psychiatric evaluations of patients. The evaluation metric used was a variation on Mean Absolute Error, to take into account the ordinal data type of the RDoc framework by which the documents were annotated. The top ten systems submitted by researcher produced results ranging from 0.863 to 0.801, with the organisers noting that the task was relatively easy to solve, with a confusion

matrix analysis revealing that classifiers often mistook the severity of neighbouring classes [163].

- Novel data use - an open ended task without a specific objective, wherein a set of mental health records were made available to the research community such that they could pursue their own research questions. Here, three teams submitted their investigations, includng attempts to predict mental conditions based on a patients historical record [164], and and exploration of the link between violence and social and clinical factors [165]. A third submission by Zhang et al [166] was of particular relevance to this thesis. Here, the researchers described work wherein they attempted to use an unsupervised apporach to extract symptoms of psychosis. The methodology used here is notable due to its parallels with my own efforts in this sphere, and is explored more in chapter 5.

Text Retrieval Conference

The medical records track of the general NLP Text Retrieval Conference (TReC) ran in 2011 and 2012, and examined the effectiveness of NLP for cohort selection using clinical documents from several hospitals. Illustrative of the issues of obtaining clinical corpora, the track was forced into indefinite hiatus after 2012 due to lack of available new datasets to support further tasks [167].

Clinical TempEval

A recent initiative, Clinical TempEval, ran in 2015 and 2016 as part of the SemEval series of evalution tasks. The original task involved three teams of participants attempting to resolve temporal relationships in clinical text, to produce an accurate sequence of events, although the task itself proved controversial, due to notably low levels of inter-annotator agreement in the supplied corpus [168].

Clinical E-Science Framework

In the UK, Roberts et al [169] describe one of the few UK specific attempts at producing a corpus intended to offer an extensively annotated set of clinical notes. The paper describes at length the semantic annotation of a corpus of 300 cancer documents for the Clinical E-Science Framework (CLEF[3]) project, noting that practical application of NLP in the clinical domain is highly dependent on annotated corpora. Although the CLEF corpus

[3]not to be confused with the Cross-Language Evaluation Forum, another group that organises generic NLP shared tassks - sometimes with a medical track.

was originally developed with specific IE outcomes in mind, the thorough consideration of the annotation process was developed such that it might be reused in the evaluation of generic clinical NLP subtasks on UK specific data. A rough estimate of the time taken to have two annotators complete the annotation process on a single document (with consensus resolution) was reportedly in the region of 1.5 hours. This is considered a major limitation in developing a corpus resource of sufficient size to allow extensive use for its stated aim of the evaluation of a multiple NLP tasks. Unfortunately, the corpus is no longer available after funding for the project ceased (Angus Roberts, personal communication).

While difficult to organise, such tasks have proven to be a useful way of coordinating the research efforts of the clinical NLP community and are valuable in producing recommendations as to the best performing methodologies for a variety of problems.

Additional Corpora

In addition to the clinical data made available via organised shared tasks, several other resources have been produced to facilitate clinical NLP development. The PhenoCHF corpus emerged contains annotated documents from both the biomedical literature and EHRs, respectively sourced from the Pubmed Central Open Access subset, and the 2008 I2B2 shared task for identification of obesity and its co-morbidities. It contains entity mention annotations to facilitate NER [170], normalisation annotations for work relating to mapping free text mentions to UMLS Metathesaurus concepts [171] and relation annotations to facilitate relationship extraction [172].

A plethora of corpora have been produced from documents sourced from the biomedical/life science literature, to facilitate shared tasks for literature mining much in the same vein as those described above. One of the most influential projects in this domain was the Genia project. The Genia project was funded from 1998 to 2012, and primarily resulted in the creation of the Genia corpus of 1 999 Medline abstracts annotated with part-of-speech, phrase structure, entities, events, relation and coreference information. the project also organised the BioNLP shared task series, that focussed on a range of tasks relevant to biomedical literature. In addition to general purpose NLP activities (such as NER), this series addressed some unusual questions particularly relevant to scientific literature, such as speculation recognition.

Przbyla et al offer a detailed discussion of further biomedical literature resources [173], in addition to tools and frameworks currently employed for such material.

2.3.8 Influential Clinical NLP Systems

Growing interest in clinical text has lead to a proliferation of publications that describe attempts to generalise NLP systems for IE. Although too many unique systems and methodologies exist to consider them all, I describe here some of the more influential developments.

Medlee [174] was one of the first attempts at a 'broad-spectrum' clinical IE system. Originally described in 1994, it was a predominantly rules based system that recognised many features of the clinical sublanguage that are widely acknowledged today, such as negation and lexical normalisation. It was validated in the context of radiology reports, with a reported 87% precision and 85% recall in the task of identifying whether patients were suffering from one or more diseases from a list of four. Its development continued after publication, with an updated position paper released in 2000 [175]. Here, problems of generalisation between domains were identified, noting that the creation and development of new rules were the main barriers to adapting the system outside of radiology. Sporadic applications of the system appeared in the literature for several years afterwards, each requiring adaption of the underlying rules [176, 177].

One of the biggest challenges to developing a general clinical IE system is offering coverage over the vast number of medical entities that exist. Many broad spectrum NLP systems seek to use standardised clinical nomenclatures or ontologies as lexical resources for this process. MetaMap, originally published in 2001 [178], attempts to offer coverage of the UMLS Metathesaurus. The project was originally devised to support accurate indexing in IR for the NLMs vast library of biomedical literature. However, it has since been additionally purposed for clinical NLP. As of 2016, it is still under active development. In addition to acting as a simple dictionary, MetaMap provides a number of additional processes including variant generation, negation detection (via NegEx) and WSD of candidate terms to mappings to UMLS concepts. Although sometimes applied directly to clinical text, MetaMap is often used as a component for more extensive clinical NLP systems. Because of this, it is hard to draw broad conclusions regarding classification performance, as such systems often employ domain adaption. The 2013 ShARe/CLEF eHealth Evaluation Lab (SHEL) was organised with the express intent of assessing the ability of general purpose clinical NLP systems to map medical concepts in clinical text directly to UMLS concepts [179]. Here, MetaMap was employed in many of the systems entered, although the general classification performance of all systems was considered poor.

Perhaps the best known broad spectrum clinical IE system is cTAKES, which has received over 1000 citations as of March 2019 [116]. Originally, cTAKES was released as a selection of the NLP subtasks described in Section 2.2 that had been optimised for clinical

text, with the intention that NLP researchers could expand upon its foundations to create higher order representations. Modules included negation detection and dictionary lookup from the SNOMED CT and RXNorm (drug names) components of UMLS. The scope of the project has since expanded to offer modules for additional entities, such as smoking status detection [180] and temporality [181]. Known issues of cTAKES include the requirement to adapt it to new domains in order to obtain optimal classification performance [182], poor scalability to handle large datasets [183, 184], and poor classification performance in handling clinical abbreviations [185].

The U-Compare platform [186], is an NLP system built upon the UIMA framework designed to facilitate the sharing and evaluation of NLP components for the biomedical domain. It was created via colalboration between the University of Tokyo, the University of Colorado School of Medicine and the National Centre for Text Mining at the University of Manchester, in recognition of the need for the NLP community to integrate the products of research into their own NLP pipelines. This platform was later superseded by Argo [187], which provided additional features to support annotation work and ease of use. As of March 2019, the Argo platform has been cited over 21 times, in manuscripts descibing both EHR and biomedical literature use cases.

Finally, MedLex [188] is an attempt at creating a semantic lexicon for EHR data via a set of heuristics. The purpose of such a lexicon is to map words and phrases as they appear in the natural language of the EHR to concrete entities as represented by standard terminologies such as SNOMED-CT. For example, the phrase as written in an EHR, "the patient had a pain in his chest" might be mapped to the snomed concept for 'chest pain' via such a lexicon. One important finding of this work was the finding that only a portion of UMLS entities were detected within the clinical corpus of the Mayo Clinic that was evaluated. My own efforts at investigating this phenomena with regard to mental health records and SNOMED-CT are described in chapter 5.

2.4 The Clinical Records Interactive Search System at The South London and Maudsley NHS Trust

The South London and Maudsley NHS Trust (SLAM) is the largest mental health organization in Europe, and is a virtual monopoly provider of mental health services to 1.2 million individuals within its geographical catchment area (Lambeth, Southwark, Lewisham and Croydon boroughs in south London). In 2007-08, funding from the British National Institute for Health Research supported the development of the Clinical Record Interactive Search (CRIS) database. As the data contained within CRIS is used extensively in this

thesis, I will describe it in some detail below.

The CRIS system was built with the specific focus of providing researchers with secure access to pseudonymised and de-identified mental health clinical records, freely accessible via its distinctive, patient-led information governance model [64]. During the course of its development, it became apparent that, as observed in many EHR systems, the use of structured data input was intermittent at best, and a vast quantity of textual data was present in the Trust's EHR system. Figure 2.6 describes this trend, where it can also be observed that the use of structured fields also show are marked decline in usage around 2009-10. Inquiries with Trust sources suggest that such a decline might be due to unsustainable management initiatives to encourage clinicians to use structured inputs.

As of August 2016, CRIS houses over 230,000 de-identified patient records, which in turn represent over 20 million free text documents. The CRIS system continues to grow at a rate of approximately 170,000 free text documents per month. However, the value of the CRIS data asset would be limited unless text can be harnessed in some way. Two key decisions were made. First, to implement a search engine to ease navigation and facilitate IR around the data. Second, to invest in NLP capacity to develop IE methodologies, thus producing structured data in a way that would make it amenable to its stated aim of supporting mental health research. This triggered a number of collaborations with various national and international interest groups, as well as the recruitment of PhD students with this objective in mind. The following section contains modified excepts that I contributed to a wider CRIS position paper [1] (See Appendix A for the full paper, describing the wider CRIS project).

2.4.1 Excerpts from the CRIS position paper

2.4.2 NLP in CRIS

In the CRIS project, NLP techniques have been evaluated and applied for extracting knowledge from unstructured text data. For our purposes, the key NLP technique has been information extraction (IE) where unstructured text is converted into structured tables [189]. Such methods promise massive reductions in the time resource required by researchers to unlock information held in clinical notes that in turn may be connected to other parts of the structured record. It was therefore decided, early in the postdevelopment phase [of the CRIS technology], to implement a text-mining capability in CRIS. This was to be generic, in that information to be extracted could not necessarily be foreseen in advance of the design of individual research studies. General Architecture for Text Engineering (GATE) was chosen as the core NLP infrastructure for CRIS [190,191]. GATE is a widely

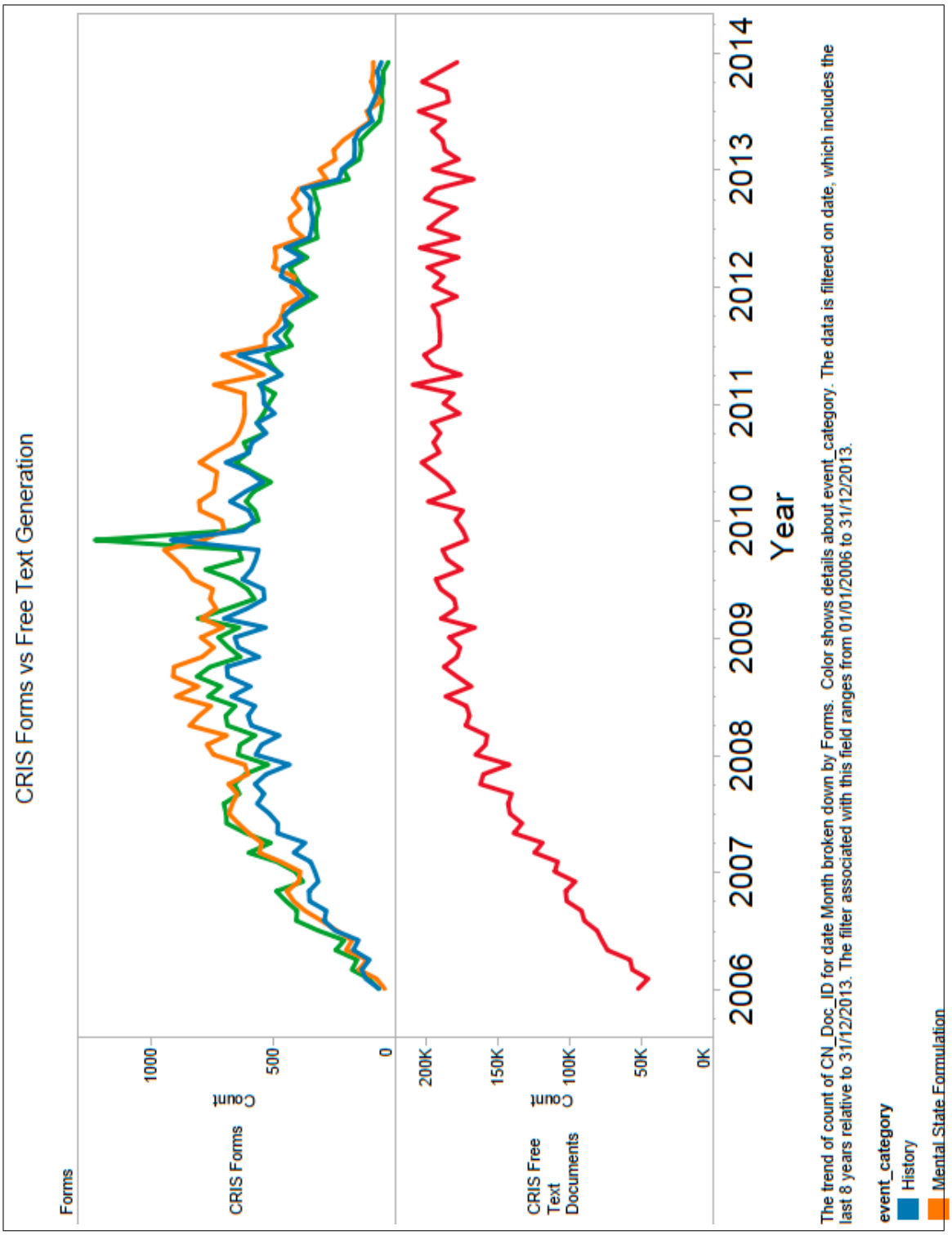


Figure 2.6: Month on month counts of the use of all freetext in the CRIS system, vs the usage of three types of structured form input, 2006-2014. Here, structured input is shown to peak around 2009-2010, and then experience a long term decline, whereas free text usage is seen to be stable.

used suite of open source software for text engineering that includes a workbench for developing applications, tools for distributing those applications on different computer hardware architectures, a quality assurance suite and facilities for manual preparation of example data [190, 191]. GATE's origins are in IE and it has been widely applied in this context [192, 193]. GATE includes a flexible architecture for IE and text mining, a large set of pluggable text processing components, and graphical tools for organising those components into new applications. The GATE suite also includes tools for text-mining workflow, distributed processing and visualisation. A variety of text processing tools and document formats may be plugged into this architecture, with individual tools being chained together into processing 'pipelines', and documents processed in series through these pipelines.

Two distinct shallow language processing methodologies have been adopted for CRIS development, in collaboration with University of Sheffield Department of Computer Science. The first may be described as rule-based pattern matching of key concepts. Sentences are first processed to find and create annotations based on simple surface linguistic information (such as words, sentences, etc). This step is then followed by the process of finding concept-specific keywords, which are used to recognise likely sentences of importance to the IE task. For example, in an application to determine the smoking status of a patient implied by texts, such a dictionary might list the terms of common tobacco products and activities—'cigarette', 'smoker', etc. Finally, a set of patterns specific to the text-mining task are run over the previously generated annotations in order to create a final annotation containing all of the information required in a readily extractable format. The challenge of the pattern matching approach is that it is knowledge intensive. A successful series of patterns need to be developed in relation to a specific IE task (eg, to extract medications, educational level or particular test results). They have to be built manually by GATE users with language engineering skills, using definitions agreed with clinicians and epidemiologists. A sample of the output from an initial prototype application is then corrected by a clinician or epidemiologist, which in turn is used to stimulate discussion about requirements and to provide a basis for multiple iterations of development until classification performance requirements are met. An advantage of this IE approach is that it also allows researchers to combine information available from open text and structured fields available in CRIS, through SQL, thus combining multiple sources of information. At the postprocessing stage, we can further apply specific filtering criteria to data extraction, such as frequency and length of prescribing and number of concomitant drugs, thus identifying more complex patterns in the text, such as antipsychotic medication profiles (ie, antipsychotic polypharmacy) [15].

Because of the lengthy development cycles of building shallow parsing algorithms, a second IE methodology has also been evaluated. Here, support vector machines (SVMs) are used to rapidly achieve respectable results for certain types of IE problem. A SVM is a machine-learning technique where the intention is to represent instances of text as vectors in high dimensional space. With a training set of instances labelled as indicative of a desired class, the SVM implementation in GATE generates a hyperplane which can in turn be used to classify unseen instances pertaining to the described class in the training set. In practice, this primarily uses a technique known as 'bag of words', where the occurrence of single words within a sentence is the principal currency used to distinguish the various classes. The first part of the model construction requires an expert (eg, clinician) to review a set of documents and label sentences which are relevant to the concept in question, in much the same way that they might signal to a language engineer the relevance of a given sentence for a pattern-based approach. The combination of labelled and unlabelled sentences forms the training data, from which the SVM learns the classification function. This model is then applied to unseen data, and the model quality assessed by human review. If required, further training data can be supplied, which may involve an active learning-inspired approach. A limitation with SVMs applied in CRIS has been that they have limited suitability for complex data extraction problems; however, in scenarios where the assertion to be extracted is simple and tend to be restricted to a concise set of clinical language, classification performance has been found to be very good and IE applications with immediate utility can be rapidly developed [20]. The TextHunter program was designed specifically to aid the process of clinical text annotation in CRIS, providing an easy-to-use interface for annotators with a focus on the sentence containing the word(s) of interest and immediately proximal text and functionality for rapid coding into discrete groups, typically comprising the following: (1) positive (ie, implying that the construct is present); (2) negative (ie, a statement indicating that the construct is absent); and (3) irrelevant text [3]. Additional TextHunter functionality includes platforms for interannotator agreement testing, and the creation of gold standard and test annotation sets.

Whether rules-based or machine-learning approaches are used, separate training and test data sets are constructed. Standard metrics for evaluating IE application classification performance in the test data sets, at the level of the individual text annotation, comprise precision (equivalent to positive predictive value; the proportion of IE application 'hits' which are found to identify the genuine construct) and recall (equivalent to sensitivity; the proportions of instances of the genuine construct which are identified by the application). Employing text mining within the CRIS data set has involved a trade-off between the two.

However, the longitudinal nature of EHR data means that there are generally multiple opportunities for an NLP application to capture a piece of information; therefore, suboptimal recall can be compensated for and the focus has been on maximising precision. For the purpose of precision and recall testing, there are two reportable outcomes. The first is 'annotation level', which is carried out across randomly selected documents and is an indicator of the base level of classification performance of the application. This figure is useful for developmental purposes, or, in the case of simple concepts that do not require postprocessing, for estimating the final classification performance of the algorithm. The second type of precision and recall are 'currency level', measuring classification performance after postprocessing.

2.4.3 Performance of NLP applications

Classification erformances of IE applications to date are summarised for CRIS as a whole, supplementary to more detailed publications on some of these [12,194–196]. The first NLP IE application to be developed was for the MMSE, a commonly used 0–30-point assessment of global cognitive function. The objective of the application was to ascertain both the numerator and denominator scores (because denominator scores of less than 30 are used where some items cannot be attempted because of, eg, sensory impairment), as well as the date implied for the assessment (because clinical text fields commonly refer to previous as well as current scores). Further rules for application postprocessing were that only MMSE scores with denominators over 25 were included (because scores below that level imply substantial missing data and a scale that was probably incompletely administered), and scores were excluded if two different numerators were assigned to the same date [196]. The application for educational attainment sought to ascertain the numeric value associated with text commenting on school leaving age, whether the age itself or the year, and the application for 'living alone' simply sought to identify that phrase or equivalents applied to the patient. In developing the smoking application, authors extracted information from open-text fields, classifying patients as either 'currently smoking', 'past smoker' or 'has never smoked', with smoking of substances other than tobacco (eg, marijuana/cannabis and cocaine) specifically excluded [194]. The methodology used an iterative process of manual 'gold standard' annotation of free-text documents, followed by comparison with the results generated by the application at each development stage, with analysis of this comparison feeding further development of the rules. The application for 'diagnosis' sought simply to extract any text strings associated with a diagnosis statement in order to supplement the existing structured (International Classification of Diseases (ICD)-10) fields. Its classification performance was evaluated formally in a random sample of 75 documents

for 'vascular dementia' [195], but is recommended for individual further evaluation in other conditions. The application for ascertaining pharmacotherapy was developed using a gazetteer of generic and commercial names for all medications in UK use in order to ascertain instances where the patient was reported as receiving these, with supplementary rules for ascertaining recorded dose, frequency/timing and starting/stopping statements. Its precision was first tested for clozapine receipt against a manual search of 279 documents, and recall was ascertained on a random set of 200 documents containing the word clozapine and scrutinised to ascertain an actual prescription [12]. Finally, the validity of this application was recently further evaluated for six antipsychotic agents (amisulpiride, flupentixol, haloperidol, olanzapine, risperidone, zuclopenthixol) on instance level (ie, specific mentions in the text at individual points in time). To estimate precision and recall, the authors examined a subset of 20 patients for each medication, totalling 120 patients (the instances of antipsychotic prescribing varied from 328 to 1150 instances by antipsychotic agent) by running the NLP application over the set of unseen documents and comparing the results to the manual coding of the same data set [15]. For all evaluations, an F-statistic was additionally calculated, representing the harmonic mean of precision and recall, and defined as: F=2*(precision*recall/(precision+recall)). As with the diagnosis application, further bespoke validation of the pharmacotherapy application is recommended for new medications or classes. Classification performance data are summarised for NLP IE applications in table 2.6, and table 2.7 describes the resulting additional structured data points generated across CRIS using these applications.

As displayed in table 2.7, the development of NLP IE applications to date has resulted in a very substantial expansion in data fields available for analysis within the SLaM BRC Case Register and in the ability to construct longitudinal data sets with repeated measures (as illustrated for MMSE score trajectories before and after initiation of dementia treatment) [197]. With increasing use of EHRs, we believe that NLP techniques have an important role to play, whether derived metadata are to be used for research or to enhance the quality of the clinical record. This is particularly pertinent for mental health records where text fields are often substantial and contain some of the most important clinical information. However, although its potential is substantial, it is important to bear in mind that there may be limits in the usefulness of NLP in EHR-sourced data resources, because of the high degree of variability in clinical text. As well as the well-recognised challenges of non-grammatical sentences, misspellings, idiosyncratic abbreviations and jargon, there are more complex issues to deal with such as the establishment of temporality (eg, timing of events described in long case summaries), the classification of documents and within-document text domains (eg, sections of the history or mental state assessment), and the

Application Name	Construct sought	Number of patients tested	Precision	Recall	F-statistic
Smoking	Is the patient a current smoker?	100	0.93	0.58	0.72
Clozapine-current use	Is the patient currently using clozapine (within 3?months)?	Precision: 279, recall: 200	0.96	0.92	0.94
Clozapine-ever used	Has the patient used clozapine in the past?	Precision: 279, recall: 200	0.99	0.92	0.95
Diagnosis	What text accompanies a statement about diagnosis?	75	0.99	0.98	0.99
MMSE	What MMSE score did the patient attain on a given date?	100	0.97	0.98	0.97
Education	What age did a patient leave school?	Precision: 100, recall: 115	0.95	0.59	0.73
Living alone	Is the patient living alone?	100	0.93	0.99	0.96
Amisulpride	Is the patient currently using amisulpride?	20 patients with 619 instances	0.97	0.61	0.75
Flupentixol	Is the patient currently using flupentixol?	20 patients with 328 instances	0.94	0.77	0.85
Haloperidol	Is the patient currently using haloperidol?	20 patients with 747 instances	0.94	0.57	0.71
Olanzapine	Is the patient currently using olanzapine?	20 patients with 1150 instances	0.95	0.69	0.8
Risperidone	Is the patient currently using risperidone?	20 patients with 737 instances	0.95	0.64	0.76
Zuclopenthixol	Is the patient currently using zuclopenthixol?	20 patients with 390 instances	0.97	0.68	0.8

Table 2.6: Classification performance of natural language processing information extraction applications developed to date in the SLaM BRC Case Register. From Perera et al. [1]

Application Name	Total Number of Instances Generated	Number of patients withat least one instance generated
MMSE	107 384	24 705
Diagnosis	615 237	78 851
Smoking	670 053	52 700
Education	181 905	51 665
Medication (selected)		
Olanzapine	371 754	25 697
Citalopram	144 072	24 363
Mirtazapine	135 309	23 710
Risperidone	240 068	22 046
Zopiclone	129 488	20 712
Diazepam	129 409	17 841
Lorazepam	119 357	15 637
Fluoxetine	96 258	15 527
Sertraline	95 381	13 600
Promethazine	112 256	12 861
Clonazepam	111 279	9679
Quetiapine	98 509	9503
Aripiprazole	90 866	8737
Haloperidol	53 936	7591
Amisulpride	58 751	6759
Methadone	128 132	6385
Flupentixol	25 576	5248
Clozapine	111 170	4364
Zuclopenthixol	18 099	3093

Table 2.7: Summary of number of annotations generated from NLP applications in the SLaM BRC Case Register. From Perera et al. [1]

development of standard ontologies, not to mention the challenges of translation and harmonisation across languages. An important decision in NLP application development at the outset is the intended use case. For instance, clinical decision support systems are likely to require high accuracy at the individual patient level, whereas observational epidemiology, which often makes use of aggregate statistics over large populations can often tolerate a larger degree of error.

2.5 Conclusion

The purpose of this chapter has been to review some of the fundamental tasks and validation techniques commonly found in IE applications, and describe the additional challenges in applying IE techniques to clinical data. I have outlined some of the global initiatives to foster a community of clinical NLP research, and described some of the influential clinical NLP systems that have emerged from the academic domain. I have introduced the concepts of rules-based and ML approaches to IE, and highlighted some of the advantages and disadvantages of each approach. Additionally, I have introduced the CRIS project and presented the results of validation work of several NLP applications that have been developed to supplement the raw CRIS data. Although effective, the development of such applications was associated with a high financial investment, owing to the high staffing costs of language engineers, the complexities of intra-institutional collaboration on clinical data and corpus development. Such issues present a bottleneck for large scale IE efforts, suggesting methodological improvements are worthy of investigation. This is the principal topic of Chapter 4.

Chapter 3

Motivation for Thesis

3.1 'Off the Shelf' Clinical NLP?

In order for a scientific method to generalise across geographic and temporal boundaries, the method has to tolerate a certain amount of error within its parameters and still retain reproducibility. To draw an example from the realm of bioinformatics, the use of microarray methods in gene expression studies suffered significant criticism during its formative years owing to the high degree of variability in replication studies, owing to the lack of standard operating procedures and widely accepted best practice. Recognising the dangers of the field losing credibility, the international community adopted significant measures to ensure the survival of the field, such as the initiation of the MAQC project [198] and journal guidelines requiring all papers made their datasets publicly available to ensure independent validation of a studies findings were possible.

Such actions proved to be successful and the technique has since continued to enjoy widespread usage [199]. However, this has only been possible because the raw inputs for microarray experimentation, RNA, has been a consistent feature throughout the history of it's usage. RNA from a zebrafish has little difference from RNA in modern Homo sapiens, nor indeed from a 2.8 million year old Homo habilis. The implication for microarray analysis is that the rate of RNA evolution, or to put it another way, the error between such inputs is sufficiently marginal that it has been possible to generalise microarray methods across them.

To contrast this with the evolution of human language, the ISO 639-3 standard suggests there are approximately 7 097 'living' languages world wide, not including extinct languages, sub-languages and other ill defined language like constructs. Within the English language alone, there are a vast number of 'mutually comprehensible' dialects [200] within English speaking communities, with an estimated 30 dialects in the British Isles alone. Language diversifies across geographic localities, and, as the field of evolutionary

linguistics will ascribe, cannot be described as temporally static. The clinical sublanguage is a technical domain, and in theory should show less variability due to its focus around a tightly defined subject matter. However, abundant issues of generalisability in clinical NLP systems suggests it is not immune to similar idiosyncrasies. This is potentially due to the fact that no two EHR systems are alike. Aside from a diverse number of EHR system providers, localised cultural, technical and political influences can contribute to potential biases in how unstructured data are recorded in EHR systems. Further, many authorship styles/document classes are known to exist within EHR systems, each requiring their own consideration [201]. One might draw comparisons with the efforts of The International Health Terminology Standards Development Organisation (distributors of SNOMED CT) with that of the MAQC project to standardise clinical terminology. However, empirically observed adoption rates of SNOMED CT within healthcare suggest that such efforts have been only partially successful [202–204]. Without the ability to draw raw data from a representative sample of the population, addressing issues of generalisability becomes a question of misaligned expectation.

What is the goal of clinical IE? Is it the linguistic study of clinical language as an end unto itself, or an intermediate process to serve business and clinical research needs? Such a question is important when considering the current trajectory of clinical NLP research output. The majority of papers that profile clinical NLP work generally follow a similar format. First, the development of an IE pipeline is described in detail, followed by an evaluation on real world data (generally from a hosting healthcare organisation who will benefit from the systems outputs) or via head-to-head comparisons with other pipelines via the shared task concept previously described.

Many authors cite the perceived benefits of clinical IE in the domains of cohort identification/phenotyping, clinical decision support and other aspects of translational medicine requiring the secondary use of clinical data. Yet a substantial fraction of the research effort in clinical IE concerns the construction of systems that are evaluated in a limited setting. The vast majority of modern research utilises hybrid or ML type methodologies in pursuit of ever higher classification performance in such limited evaluations. To date, the most effective ML approaches require large corpora of context specific annotated data to reach moderate to high levels of classification performance. Much progress has been made in the availability of corpora for clinical NLP via the activities of shared task organisers and EHR research enclaves. However, the fact remains that there are few such resources available relative to the scale and diversity of the world's clinical narrative, and access to such resources is still extremely restrictive. Specifically regarding the purpose of clinical NLP shared tasks, such activities serve as useful bellweathers as to technical progress in

the field, and offer evidence of current thinking regarding best practice when considering an NLP task. Nevertheless, such corpora cannot be considered sufficiently diverse to address the question of how the model will perform on other healthcare datasets, without substantial modification. Hence, while the development of clinical IE systems might favour 'adaptability', little progress has been made regarding 'generalisability' (in the sense that a fully fledged clinical NLP system can be plugged in to a wider range of clinical document systems and good results can be obtained).

To cite a specific example, one might regard the subject of negation detection. The continuing popularity of the NegEx algorithm might be considered a function of the constraints of the clinical sublanguage. Since there are generally few observed ways to express negation in clinical text, the resulting NLP problem is also markedly less difficult than detecting equivalent negations in standard grammar. Widespread success with negation problems had (for a time), prompted a declining interest in tackling negation, suggesting a growing consensus that negation is a 'solved problem' in a generalised sense. However, Wu et al [205] and Koeling et al [206] argued that while negation systems were easy to 'optimise' (in the sense of adapting a system to a new domain), it was inappropriate to consider the concept of negation 'solved' in the sense that a given negation solution would work 'off-the-shelf'. Wu and colleagues observed that the best classification performance in negation problems resulted from systems adapted to a specific corpus, and when applied directly to a new corpus without modification, marked drops in classification performance were observed. In conclusion, Wu et al note that the most reliable means to tackle negation problems is to apply domain adaptation techniques or modify rule sets to a target domain. In addition, recent work from Zamaraeva et al [207] has revisited the negation problem in clinical text by making use of a precision grammar in feature extraction, with favourable results compared to NegEx.

One might question whether a time will arrive when sufficient clinical corpora will be available to meet the challenges of generalisability, given the often uncompromising approach many organisations take to protect healthcare data. Such a situation presents a critical issue for the goal of generalisability, hampering efforts for data sharing, collaboration and independent validation of clinical NLP methods. While efforts to develop and improve the classification performance of clinical NLP subtasks are highly valued activities, the weight of public attitudes towards privacy considerations as evidenced by the scarcity of publicly available clinical text corpora suggests that the widespread sharing of clinical data will not be a realistic objective in the short to medium term, and therefore one might assume progress on general purpose clinical NLP systems will continue to struggle.

3.2 Realising the Potential of ML methods in the Clinical Setting

If generalisability is a grand challenge in clinical IE, how can the field deliver real world value in the interim? In acknowledgement of promising advances in the field of IE, Nadkarni et al [208] suggested that the commoditisation of clinical NLP was overdue. However, an idealogical disconnect exists between academic and industrial doctrine. In many situations, rule based approaches are considered in academic circles to be too time-consuming and expensive to produce in high volume [128], prompting the growth in popularity of ML methods in academic circles. A recent comparison of academic and industry approaches for IE problems found that commercial IE products overwhelmingly favoured rule based approaches over ML [143]. In spite of ten years of academic focus on ML IE methodologies, outputs have largely failed to influence commercial realisations. The divergence in practices serves to illustrate the conflict between the perceived benefits of clinical NLP theory and the practical application of IE to solve real world problems. Although both industry and academia are in agreement about the benefits and drawbacks of each approach, the authors suggested that there is a misaligned value proposition of the role of IE between academia and industry.

One group of factors concerns the differing perceptions of the benefits of IE. First, attempts at generic clinical IE systems often fail to identify the differential business value of the enormous range of clinical entities. When comparing the results of different clinical NLP systems, a corpus of clinical data produced for a theoretical shared task weighs the contribution of the entities "broken toe" and "pancreatic cancer" evenly in terms of producing precision and recall metrics. However, the business implications of a patient with pancreatic cancer are clearly more severe. Therefore, while a large number of medical entities exist, their business value is heavily skewed. Producing a small set of high classification performance rules targeting high value concepts is a more cost effective solution in terms of addressing primary business requirements than the implementation of theoretical best practice. A separate consideration is the perennial complaint against rule based systems is that they are time-consuming and monotonous to develop, and are primitive compared to the developments in the ML domain. While aspects of this are undoubtedly true, this argument fails to acknowledge the costs of developing extensively annotated corpora to adapt ML approaches to a specific domain. Such costs come in many forms, such as developing appropriate guidelines, training annotators and managing the annotation process and associated staffing costs.

Chiticariu et al further comment on the dangers of generalising the results of a small

sample of entities over a much larger set:

"*... In real-world applications, the output of extraction is often the input to a larger process, and it is the quality of the larger process that drives business value. This quality may derive from an aspect of extracted output that is only loosely correlated with overall precision and recall. For example, does extracted sentiment, when broken down and aggregated by product, produce an unbiased estimate of average sentiment polarity for each product?* "

In a clinical context, this might be interpreted as extrapolating the generalisability of a clinical IE system over the one third of a million concepts in SNOMED CT, based on results calculated on a gold standard corpus containing just a few hundred. The potential implication for such systems is that they serve all needs 'moderately', delivering vast amounts of mediocre quality data that are surplus to core business requirements.

Second, the authors suggest that ML approaches are often ill-equipped to deal with the protean, ill-defined nature of business data (which would include clinical data, in our context). The concept of developing universal annotation guidelines for clinical data to facilitate cross-domain validation of tools implies that there are universal 'truths' in how clinical documents are authored. This is evidenced by the ubiquitous use of precision and recall statistics in NER and other IE tasks, suggesting such entities can be objectively defined. While this may be true for many relevant concepts in the clinical sublanguage (such as whether a clinician prescribes a medicine or not [209]), much of clinical text concerns investigatory reporting, rather than capturing absolute truths [83]. Here, the interpretation of clinical language is determined just as much by the context of the reader as it is the intentions of the author. For instance, it seems dangerous to assume that guidelines produced in California regarding how some facet of the clinical domain is expressed should also apply to the bespoke practices, cultures and clinical systems operating in a remote part of Scotland. Such a situation is further compounded where controversy exists regarding disease definitions, as is the case in many areas of mental health (see chapter 4 for details). A safer approach is to assume localised knowledge of linguistic context by subject matter experts holds more value than the often opaque constructs of global origin.

Third, a feature of commercial NLP tools is that the companies behind them recognise that their commercial value is limited by the size of their potential user base. To this end, the software houses behind them invest heavily in the user experience to make them as accessible as possible to a wide variety of non-technical users. In doing so, they enable subject matter experts to interface directly with the raw data, rapidly iterating through

a series of decisions in order to produce an output according to their understanding of the domain [210]. By eliminating the need for an NLP 'middleman', the risk of miscommunicating the business objective of an IE task is reduced and the overall efficiency of the task is increased.

By contrast, references to user experience of clinical NLP systems resulting from academic sources are sparse. In recognition of the importance of this problem, the 2014 I2B2 challenge issued a specific track to assess the general usability of a range of academic software. The results suggested the assessed systems were "extremely difficult to use and understand" [211]. Concretely, applications arising from research labs tend to suffer from poor, ill-maintained documentation and are rarely built in consideration of how they might interface with other systems or address issues of scalability. For example, official statistics for MetaMap suggest substantial system requirements. Even if these are met, the system will take about 40 minutes to process a mere 1.23MB of text [212].

Finally, the maintainence of NLP systems in production environments is rarely discussed, in spite of the real world consequences of failing to manage such issues. For instance, the dynamic nature of the clinical sublanguage can be empiracally observed within the context of the CRIS project in the case of an application to extract Mini Mental State Exam (MMSE) scores. Here, a rules-based application was developed to extract MMSE scores from clinical notes by a team at the University of Sheffield, and had worked to a certain standard for a number of years [1]. However, towards the end of 2012, the extraction rates of the application experienced a marked drop in the number of instances it was producing (Figure 3.1). An investigation showed that Psychological Assessment Resources Inc. had exerted copyright over the term "Mini Mental State Exam" and its abbreviations. Consequently, the Trust started to employ a similar version of the test called the 'SMMSE'. Such an action affected the use of language in the Trust (presumably to avoid potential litigation). Rather than referring to the test as 'MMSE', staff began referring to the test as the 'SMMSE'. The rules-based nature of the application meant that it was possible to modify the NER component with a few extra dictionary terms, and normal service was restored. However, had this component been based on an ML model, it may have been necessary to undergo potentially costly retraining of the model with additional data, or apply a rules based preprocessing component to normalise the 'SMMSE' and 'MMSE' terms.

Although the arguments of Chiticariu et al principally concern the divide between rule based and ML approaches, such an argument can be abstracted to suggest that the issue is not with ML per se, but the manner in which current research approaches to IE problems do not consider the full picture of how research outputs might be utilised in

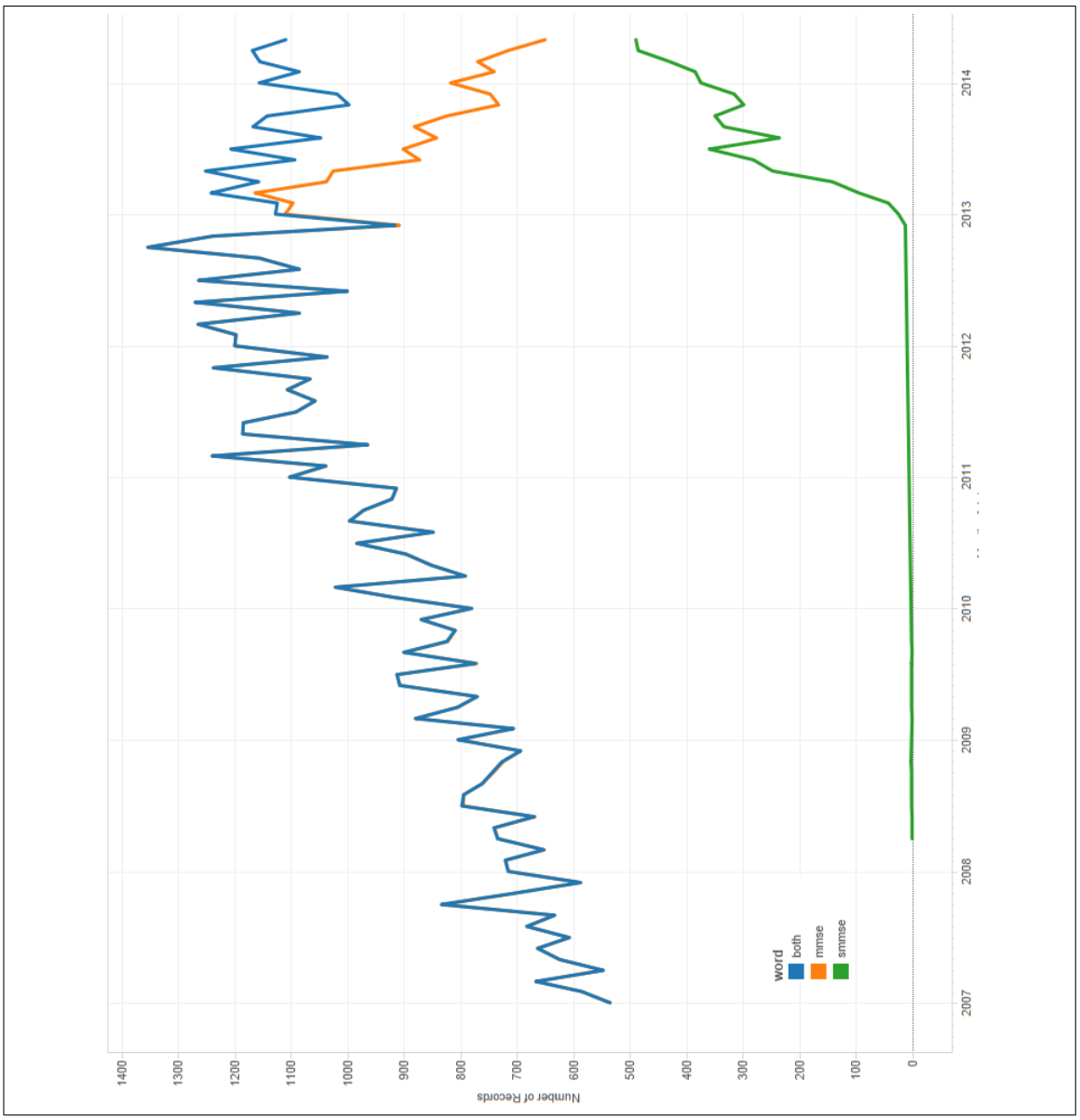


Figure 3.1: Counts of documents in CRIS mentioning MMSE, SMMSE and both terms, 2007 - 2014.

real world applications. Much progress has been made in identifying ML techniques as best practice to address IE problems within the NLP domain. However, applied clinical NLP necessitates an appreciation of the context in which a system is to be used, both in terms of the scope of the problem and the technical and resource constraints of the environment. One might argue that it is not the role of clinical NLP research to address such practical considerations, yet to not do so risks widening the gulf between pure research and practical application. How then, can the best theoretical practice be brought back into line with practical considerations? A central theme of this thesis is that in order to realise the aspirations of secondary use of clinical text, clinical NLP needs to apply as much weight to pragmatic issues as it does to theoretical ones. We 'know' there are a set of approaches and algorithms that will give good classification performance on certain types of problems. We also 'know' that the difference in classification performance between well known algorithms typically employed on such problems tends to be marginal. Therefore, those seeking to enjoy the practical benefits of IE for downstream use cases might be minded to find a solution that involves the fewest hurdles to jump between a corpora of raw clinical text and the structured outputs they desire.

In order for healthcare organisations to take full advantage of the data organisation capabilities of NLP, future directions should avoid the dogma of building and validating models of clinical language for one-off projects and competitions, and focus on the portability and ease of use of the best methods within specific institutional contexts. Conceptualising what constitutes 'the best methods' should move beyond simple precision and recall statistics upon standard corpora, to address real world application issues such as speed/scalability, governance issues created by the reliance of external resources, maintainability in production and ease of use by non-specialist staff.

3.3 Scope of the thesis

In this thesis, I explore some of the obstacles described above, and how addressing them has benefited downstream clinical researchers in their work. Over the following chapters, I focus on annotation work, model building, evaluation, vocabulary building and scaling solutions within (and occasionally outside of) the domain of SMI.

3.4 Conclusion

While the field of NLP is considerably older than that of microarray analysis, the nature of language is subject to constant change, driven by socio-political elements; advancements in research requiring expansions of semantics; and the phenomena of language change.

The concept of applied clinical NLP under such circumstances might therefore seem to depend on the ability to rapidly adapt to changing requirements and practical settings. Such aspects provide little comfort for clinical researchers seeking to utilise NLP for practical purposes. Methods that seek to bridge the gap between cutting edge theoretical classification performance and applications to real world use cases are likely to be received favourably for the real world use cases of clinical NLP.

Chapter 4

Streamlining Information Extraction Methodologies for Mental Health Symptomatology

4.1 Overview

In the previous chapter, I made an argument that general purpose clinical NLP systems are not feasible at the current time, and that the the effective use of ML methods in clinical NLP are limited by the dependency of having trained ML practitioners available. In this chapter, I present work that attempts to offer an alternative approach by removing some of the barriers to entry for effective ML practise by non-specialist users. The domain for this work is mental health symptomatology.

In mental health, there is an increasing emphasis on using dimensional symptom scales to define mental illness rather than discrete diagnostic categories [213–216]. As a consequence, many mental health EHR systems do not offer structured inputs to capture symptomatology. Such concepts are therefore a high value target for IE. I present three papers (two first author, one co-author):

- First, an investigation to gauge the feasibility of rules-based, ML and hybrid approaches for symptomatology extraction.
- Second, a refinement of these methods and the development of the TextHunter application to lower the technical barrier for using ML methods for IE.
- Third, the CRIS-CODE project that describes the requirement for IE of the symptomatology of Serious Mental Illness and the operation of the TextHunter application at scale to achieve widespread coverage of Serious Mental Illness Symptomatology.

4.1.1 Methods for Extracting Negative Symptomatology

A basic ML methodology was first proposed, developed and evaluated to extract a range of negative symptoms of schizophrenia. This work was co-authored with collaborators at the University of Sheffield, and presented by me at the The 9th International Conference on Recent Advances in Natural Language Processing [20] in the peer reviewed Workshop on NLP for Medicine and Biology. Genevieve Gorrell is the first author and developed the methodology. Angus Roberts supervised Genevieve and contributed to the design. Robert Stewart supervised me and provided the raw annotation of the data. I was responsible for managing the collaboration relationship between Sheffield and King's College London, including collecting/provisioning the data from the CRIS database, developing a simple annotation process, providing additional annotation work and cleaning the data for use by Genevieve.

I reproduce the paper in full below.

Finding Negative Symptoms of Schizophrenia in Patient Records

Genevieve Gorrell
The University of Sheffield
g.gorrell@sheffield.ac.uk

Angus Roberts
The University of Sheffield
a.roberts@dcs.shef.ac.uk

Richard Jackson
King's College London
Richard.G.Jackson@slam.nhs.uk

Robert Stewart
King's College London
robert.stewart@kcl.ac.uk

Abstract

This paper reports the automatic extraction of eleven negative symptoms of schizophrenia from patient medical records. The task offers a range of difficulties depending on the consistency and complexity with which mental health professionals describe each. In order to reduce the cost of system development, rapid prototypes are built with minimal adaptation and configuration of existing software, and additional training data is obtained by annotating automatically extracted symptoms for which the system has low confidence. The system was further improved by the addition of a manually engineered rule based approach. Rule-based and machine learning approaches are combined in various ways to achieve the optimal result for each symptom. Precisions in the range of 0.8 to 0.99 have been obtained.

1 Introduction

There is a large literature on information extraction (IE) from the unstructured text of medical records (see (Meystre et al., 2008) for the most recent review). Relatively little of this literature, however, is specific to psychiatric records (see (Sohn et al., 2011; Lloyd et al., 2009; Roque et al., 2011) for exceptions to this). The research presented here helps to fill this gap, reporting the extraction of schizophrenia symptomatology from free text in the case register of a large mental health unit, the South London and Maudsley NHS Trust (SLaM).

We report the extraction of negative symptoms of schizophrenia, such as poor motivation, social withdrawal and apathy. These often present in addition to more prominent, positive symptoms such as delusions and hallucinations. Negative symptoms can severely impair the quality of life of affected patients, yet existing antipsychotic medications have poor efficacy in their treatment. As negative symptoms can be measured in quantitative frameworks within a clinical environment (Kay et al., 1987; Andreasen, 1983), they have the potential to reflect the success or failure of new medical interventions, and are of widespread interest in the epidemiology of schizophrenia. The motivation for our work is to provide information on the presence of negative symptoms, for use in such quantitative measures.

SLaM covers a population of 1.1 million, being responsible for close to 100% of the mental health care contacts in four London boroughs. Approximately 225,000 records are stored in the SLaM Electronic Health Record (EHR) system, which supports an average of 35,000 patients at any one time. SLaM hosts the UK National Institute for Health Research (NIHR) Biomedical Research Center (BRC) for Mental Health. The BRC deidentifies all records in the SLaM EHR (Fernandes et al., 2013) to form the largest mental health case register in Europe, the Case Register Interactive Search (CRIS) system (Stewart et al., 2009). CRIS provides BRC epidemiologists with search facilities, via a web front end that allows standard information retrieval queries over an inverted index, and via database query languages. CRIS has been approved as an anonymized data resource for secondary analysis by Oxfordshire Research Ethics Committee C (08/H0606/71). The governance for all CRIS projects and dissemination is managed through a patient-led oversight committee.

CRIS contains both the structured information, and the unstructured free text from the SLaM

Proceedings of the Workshop on NLP for Medicine and Biology associated with RANLP 2013, pages 9–17,
Hissar, Bulgaria, 13 September 2013.

EHR. The free text consists of 18 million text field instances – a mix of correspondence and notes describing patient encounters. Much of the information of value to mental health epidemiologists is found in these free text fields. SLaM clinicians record important information in the textual portion of the record, even when facilities are provided for recording the same information in a structured format. For example, a query on the structured fields containing Mini Mental State Examination scores (MMSE, a score of cognitive ability) recently returned 5,700 instances, whereas a keyword search over the free text fields returned an additional 48,750 instances. The CRIS inverted index search system, however, cannot return the specific information of interest (the MMSE score in this case), instead returning each text field that contains a query match, in its entirety. In the case of symptomatology, as examined in this paper, symptoms are rarely recorded in structured fields, but are frequently mentioned in the unstructured text.

This problem is not unusual. (Meystre et al., 2008) note that free text is "convenient to express concepts and events" (Meystre et al., 2008), but that it is difficult for re-use in other applications, and difficult for statistical analysis. (Rosenbloom et al., 2011) have reviewed the few studies that look at the expressivity of structured clinical documentation systems compared to natural prose notes, and report that prose is more accurate, reliable and understandable. (Powsner et al., 1998) refer to structured data as freezing clinical language, and restricting what may be said. (Greenhalgh et al., 2009), referring to the free text of the paper record, say that it is tolerant of ambiguity, which supports the complexity of clinical practice. Much of medical language is hedged with ambiguity and probability, which is difficult to represent as structured data (Scott et al., 2012).

Given the presence of large quantities of valuable information in the unstructured portion of the BRC case register, and CRIS's inability to extract this information using standard information retrieval techniques, it was decided, in 2009, to implement an IE and text mining capability as a component of CRIS. This comprises tools to develop and evaluate IE applications for specific end-user requirements as they emerge, and the facility to deploy these applications on the BRC compute cluster.

Most IE applications developed by the BRC to date have used a pattern matching approach. In this, simple lexico-syntactic pre-processing and dictionary lookup of technical terms are followed by cascades of pattern matching grammars designed to find the target of extraction. These grammars are hand-written by language engineers. Previous extraction targets have included smoking status, medications, diagnosis, MMSE, level of education, and receipt of social care. Building such pattern matching grammars is often time consuming, in that it takes significant language engineer time to develop and refine grammars. In addition, the process of writing and testing grammars requires examples of the extraction target. These are provided by manual annotation, or labelling, of examples and correction of system output; a task which takes significant domain expert time.

In the case of schizophrenia, the IE applications are required to extract multiple symptoms for use in quantitative measures of the disease. The set of symptoms relevant to such quantitative measures number in the dozens. Given the cost of pattern grammar development, and the cost of manual annotation, it is impractical to develop grammars for each of the required symptoms, and such an approach would not scale up to larger numbers of symptoms and to other diseases. In addition, the cost of domain expert annotation of examples for each individual symptom is also high. The approach taken in our research aims to reduce these two costs.

In order to reduce the cost of system development, and to improve scalability to new symptoms and diseases, we build rapid prototypes, using off-the-shelf NLP and machine learning (ML) toolkits. Such toolkits, and repositories of applications built on them, are becoming increasingly popular. It has been asked (Nadkarni et al., 2011) whether such tools may be used as "commodity software" to create clinical IE applications with little or no specialist skills. In order to help answer this question, we compare the performance of our ML only prototypes to applications that combine ML and pattern matching, and to applications implemented with pattern matching alone.

The second cost considered is that of finding and labelling high quality examples of the extraction target, used to inform and test system development. To deal with this cost, we explore methods of enriching the pool of examples for labelling,

including the use of methods inspired by active learning (Settles, 2012). In active learning, potential examples of the extraction target are selected by the learning algorithm for labelling by the human annotator. The aim is to present instances which will most benefit the ML algorithm, at least human cost. This paper presents results from experiments in training data enrichment, and a simple approach to active learning, applied to symptom extraction.

The paper is organised as follows. Section 2 looks at the task domain in more detail, explaining the symptoms to be extracted, and describing the dataset. Section 3 describes the experimental method used, and the evaluation metrics. This is followed by a presentation of the results in Section 4, and a discussion of these results in Section 5. Finally, we draw some conclusions in Section 6.

2 Analysis of the Task Domain

In this section we will first introduce the concept of negative symptoms and explain what entities we are aiming to extract from the data. We will then discuss the datasets we used, and how each symptom varies in its nature and therefore difficulty.

2.1 Negative Symptoms

In the psychiatric context, negative symptoms are deficit symptoms; those that describe an absence of a behaviour or ability that would normally be present. A positive symptom would be one which is not normally present. In schizophrenia, positive symptoms might include delusions, auditory hallucinations and thought disorder. Here, we are concerned with negative symptoms of schizophrenia, in particular the following eleven, where bold font indicates the feature values we hope to extract from the data (in machine learning terms, the classes, not including the negative class). Examples illustrate something of the ways in which the symptom might be described in text. "ZZZZZ" replaces the patient name for anonymization purposes:

- **Abstract Thinking**: Does the individual show evidence of requiring particularly **concrete** conceptualizations in order to understand? Examples include; "Staff have noted ZZZZZ is very concrete in his thinking", "Thought disordered with concrete thinking", but NOT "However ZZZZZ has no concrete plans to self-harm"

- **Affect**: Is the individual's emotional response **blunted** or **flat**? Is it inappropriate to events (**abnormal**)? Alternatively, does the individual respond appropriately (**reactive**)? Examples include; "Mood: subjectively 'okay' however objectively incongruent", "Denied low mood or suicide ideation", "showed blunting of affect"

- **Apathy**: Does the individual exhibit **apathy**? Examples include; "somewhat apathetic during his engagement in tasks", "Apathy."

- **Emotional Withdrawal**: Does the individual appear **withdrawn** or **detached**? Examples include; "withdrawal from affectational and social contacts", "has been a bit withdrawn recently", NOT "socially withdrawn", which is a separate symptom, described below.

- **Eye Contact**: Does the individual make **good** eye contact, or is it **intermediate** or **poor**? Examples include; "eye contact was poor", "maintaining eye contact longer than required", "made good eye contact"

- **Motivation**: Is motivation **poor**? Examples include; "ZZZZZ struggles to become motivated.", "ZZZZZ lacks motivation.", "This is due to low motivation."

- **Mutism**: A more extreme version of poverty of speech (below), and considered a separate symptom, is the individual **mute** (but not deaf mute)? Examples include; "Was electively mute [...]", "ZZZZZ kept to himself and was mute.", NOT "ZZZZZ is deaf mute."

- **Negative Symptoms**: An umbrella term for the symptoms described here. Do we see any **negative symptom**? Examples include; "main problem seems to be negative symptoms [...]", "[...] having negative symptoms of schizophrenia."

- **Poverty of Speech**: The individual may show a deficit or **poverty** of speech, or their speech may be **abnormal** or **normal**. Examples include; "Speech: normal rate and rhythm", "speech aspontaneous", "speech

was dysarthric", "ongoing marked speech defect", "speech was coherent and not pressured"

- **Rapport**: Individual ability to form conversational rapport may be **poor** or **good**. Examples include; "we could establish a good rapport", "has built a good rapport with her carer"

- **Social Withdrawal**: Do we see indications of **social withdrawal** or not? Examples include; "long term evidence of social withdrawal", "ZZZZZ is quite socially withdrawn"

2.2 Dataset

Different symptoms vary in the challenges they pose. For example, "apathy" is almost exclusively referred to using the word "apathy" or "apathetic", and where this word appears, it is almost certainly a reference to the negative symptom of apathy, whereas concrete thinking is harder to locate because the word "concrete" appears so often in other contexts, and because concrete thinking may be referred to in less obvious ways. In the previous section, we gave some examples of negative symptom mentions that give an idea of the range of possibilities. Exemplars were unevenly distributed among medical records, with some records having several and others having none.

Due to the expertize level required for the annotation part of the task, and strict limitations on who is authorized to view the data, annotation was performed by a single psychiatrist. Data quantity was therefore limited by the amount of time the expert annotator had available for the work. For this reason, formal interannotator agreement assessment was not possible, although a second annotator did perform some consistency checking on the data. Maximizing the utility of a limited dataset therefore constituted an important part of the work.

Because many of the records do not contain any mention of the symptom in question, in order to make a perfect gold standard corpus the expert annotator would have to read a large number of potentially very lengthy documents looking for mentions that are thin on the ground. Because expert annotator time was so scarce, this was likely to lead to a much reduced corpus size, and so a compromise was arrived at whereby simple heuristics were used to select candidate mentions for the annotator to judge rather than having also to find them. For example, in abstract thinking, one heuristic used was to identify all mentions of "concrete". In some cases, the mention is irrelevant to concrete thinking, so the annotator marks it as a negative, whereas in others it is a positive mention. This means that compared with a fully annotated corpus, our data may be lower on recall, since some cases may not have been identified using the simple heuristics, though precision is most likely excellent, since all positive examples have been fully annotated by the expert. In terms of the results reported here, this compromise has little impact, since the task is defined to be replicating the expert annotations, whatever they may be. However, it might be suggested that our task is a little easier than it would have been for a fully annotated corpus, since the simple heuristics used to identify mentions would bias the task toward the easier cases. In terms of the adequacy of the result for future use cases, precision is the priority so this decision was made with end use in mind.

2.2.1 Selecting examples for training

As a further attempt to obtain more expert-annotated data, the principles of active learning were applied in order to strategically leverage annotator time on the most difficult cases and for the most difficult symptoms. Candidate mentions were extracted with full sentence context on the basis of their confidence scores, as supplied by the classifier algorithm, and presented to the annotator for judgement. Mentions were presented in reverse confidence score order, so that annotator time was prioritized on those examples where the classifier was most confused.

3 Method

Because the boundaries of a mention of a negative symptom are somewhat open to debate, due to the wide variety of ways in which psychiatric professionals may describe a negative symptom, we defined the boundaries to be sentence boundaries, thus transforming it into a sentence classification task. However, for evaluation purposes, precision, recall and F1 are used here, since observed agreement is not appropriate for an entity extraction task, giving an inflated result due to the inevitably large number of correctly classified negative examples.

Due to the requirements of the use case, our work was biased toward achieving a good preci-

sion. Future work making use of the data depends upon the results being of good quality, whereas a lower recall will only mean that a smaller proportion of the very large amount of data is available. For this reason, we aimed, where possible, to achieve precisions in the region of 0.9 or higher, even at the expense of recalls below 0.6.

Our approach was to produce a rapid prototype with a machine learning approach, and then to combine this with rule-based approaches in an attempt to improve performance. Various methods of combining the two approaches were tried. Machine learning alone was performed using support vector machines (SVMs). Two rule phases were then added, each with a separate emphasis on improving either precision or recall. The rule-based approach was then tried in the absence of a machine learning component, and in addition both overriding the ML where it disagreed and being overridden by it. Rules were created using the JAPE language (Cunningham et al., 2000). Experiments were performed using GATE (Cunningham et al., 2011; Cunningham et al., 2013), and the SVM implementation provided with GATE (Li et al., 2009).

Evaluation was performed using fivefold cross-validation, to give values for precision, recall and F1 using standard definitions. For some symptoms, active learning data were available (see Section 2.2.1) comprising a list of examples chosen for having a low confidence score on earlier versions of the system. For these symptoms, we first give a result for systems trained on the original dataset. Then, in order to evaluate the impact of this intervention, we give results for systems trained on data including the specially selected data. However, at test time, these data constitute a glut of misrepresentatively difficult examples that would have given a deflated result. We want to include these only at training time and not at test time. Therefore, the fold that contained these data in the test set was excluded from the calculation. For these symptoms, evaluation was based on the four out of five folds where the active learning data fell in the training set. The symptoms to which this applies are abstract thinking, affect, emotional withdrawal, poverty of speech and rapport.

In the next section, results are presented for these experiments. The discussion section focuses on how results varied for different symptoms, both in the approach found optimal and the result achieved, and why this might have been the case.

4 Results

Table 1 shows results for each symptom obtained using an initial "rapid prototype" support vector machine learner. Confidence threshold in all cases is 0.4 except for negative symptoms, where the confidence threshold is 0.6 to improve precision. Features used were word unigrams in the sentence in conjunction with part of speech (to distinguish for example "affect" as a noun from "affect" as a verb) as well as some key terms flagged as relevant to the domain. Longer n-grams were rejected as a feature due to the small corpus sizes and consequent risk of overfitting. A linear kernel was used. The soft margins parameter was set to 0.7, allowing some strategic misclassification in boundary selection. An uneven margins parameter was used (Li and Shawe-Taylor, 2003; Li et al., 2005) and set to 0.4, indicating that the boundary should be positioned closer to the negative data to compensate for uneven class sizes and guard against small classes being penalized for their rarity. Since the amount of data available was small, we were not able to reserve a validation set, so care was taken to select parameter values on the basis of theory rather than experimentation on the test set, although confidence thresholds were set pragmatically. Table 1 also gives the number of classes, including the negative class (recall that different symptoms have different numbers of classes), and number of training examples, which give some information about task difficulty.

As described in Section 2.2.1, active learning-style training examples were also included for symptoms where it was deemed likely to be of benefit. Table 2 provides performance statistics for these symptoms alongside the original machine learning result for comparison. In all cases, some improvement was observed, though the extent of the improvement was highly variable.

Central to our work is investigating the interplay between rule-based and machine learning approaches. Rules were prepared for most symptoms, with the intention that they should be complementary to the machine learning system, rather than a competitor. The emphasis with the rules is on coding for the common patterns in both positive and negative examples, though coding the ways in which a symptom might not be referred

Table 1: Machine Learning Only, SVM

Symptom	Classes	Training Ex.	Precision	Recall	F1
Abstract Thinking	2	118	0.615	0.899	0.731
Affect	5	103	0.949	0.691	0.8
Apathy	2	145	0.880	0.965	0.921
Emotional Withdrawal	3	118	0.688	0.815	0.746
Eye Contact	4	35	0.827	0.677	0.745
Motivation	2	259	0.878	0.531	0.662
Mutism	2	234	0.978	0.936	0.956
Negative Symptoms	2	185	0.818	0.897	0.856
Poverty of Speech	4	263	0.772	0.597	0.674
Rapport	3	139	0.775	0.693	0.731
Social Withdrawal	2	166	0.940	0.958	0.949

Table 2: Active Learning

Symptom	Ex.	Without AL-Style Examples			With AL-Style Examples			Difference
		Prec	Rec	F1	Prec	Rec	F1	
Abstract Thinking	99	0.595	0.940	0.728	0.615	0.899	0.731	0.003
Affect	200	0.947	0.529	0.679	0.949	0.691	0.8	0.121
Emotional Withdrawal	100	0.726	0.517	0.604	0.688	0.815	0.746	0.142
Poverty of Speech	62	0.721	0.515	0.601	0.772	0.597	0.674	0.073
Rapport	37	0.725	0.621	0.669	0.775	0.693	0.731	0.062

to is considerably harder. F1 results for the standalone rule-based systems where sufficiently complete are given in Table 4; however, for now, we focus on the results of our experiments in combining the two approaches, which are given in Table 3. Here, we give results for layering rules with machine learning. On the left, we see results obtained where ML first classifies the examples, then the rule-based approach overrides any ML classification it disagrees with. In this way, the rules take priority. On the right, we see results obtained where machine learning overrides any rule-based classification it disagrees with. The higher of the F1 scores is given in bold. Results suggest that the more successful system is obtained by overriding machine learning with rules rather than vice versa.

Table 4 gives a summary of the best results obtained by symptom, using all training data, including active learning instances. We focus on F1 scores only here for conciseness. The baseline machine learning result is first recapped, along with the rule-based F1 where this was sufficiently complete to stand alone. Since in all cases, overriding machine learning with rules led to the best result of the two combination experiments, we give the F1 for this, which in all cases, where available, proves the best result of all. We provide the percentage improvement generated relative to the ML baseline by the combined approach. The final column recaps the best F1 obtained for that symptom. We can clearly see from Table 4 that in all cases, the result obtained from combining approaches outperforms either of the approaches taken alone.

5 Discussion

In summary, the best results were obtained by building upon a basic SVM system with layers of rules that completed and corrected areas of weakness in the machine learning. Note that the symptoms where this approach yielded the most striking improvements tended to be those with the fewer training examples and the larger numbers of classes. In these cases, the machine learning approach is both easier to supplement using rules and easier to beat. A high performing rule-based system certainly correlates with a substantial improvement over the ML baseline; however, we

Table 3: Machine Learning Layered with Rules

Symptom	**Rules Override ML**			**ML Overrides Rules**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Abstract Thinking	0.914	0.719	**0.805**	0.935	0.652	0.768
Affect	0.931	0.827	0.876	0.931	0.827	0.876
Emotional Withdrawal	0.840	0.778	**0.808**	0.691	0.827	0.753
Eye Contact	0.88	0.852	**0.866**	0.779	0.611	0.684
Mutism	0.986	0.936	**0.960**	0.978	0.936	0.956
Negative Symptoms	0.851	0.897	**0.874**	0.818	0.897	0.856
Poverty of Speech	0.8	0.730	**0.763**	0.793	0.723	0.757
Rapport	0.839	0.868	**0.853**	0.907	0.772	0.834

Table 4: Best Result Per Symptom

Symptom	**Classes**	**Ex.**	**ML F1**	**Rules F1**	**Rules>ML F1**	**% Imp**	**Best F1**
Abstract Thinking	2	217	0.731	0.765	0.805	10%	0.805
Affect	5	303	0.800	0.820	0.876	9%	0.876
Apathy	2	145	0.921	n/a	n/a	n/a	0.921
Emotional withdrawal	3	218	0.746	0.452	0.808	8%	0.808
Eye contact	4	35	0.745	0.859	0.866	16%	0.866
Motivation	2	259	0.662	n/a	n/a	n/a	0.662
Mutism	2	234	0.956	n/a	0.960	0%	0.960
Negative Symptoms	2	185	0.856	n/a	0.874	2%	0.874
Poverty of speech	4	325	0.674	0.689	0.763	13%	0.763
Rapport	3	176	0.731	0.826	0.853	17%	0.853
Social withdrawal	2	166	0.949	n/a	n/a	n/a	0.949

do also consistently see the combined approach outperforming both the ML and rule-based approaches as taken separately. We infer that this approach is of the most value in cases where training data is scarce.

Where machine learning was removed completely, we tended to see small performance decreases, but in particular, recall was badly affected. Precision, in some cases, improved, but not by as much as recall decreased. This seems to suggest that where datasets are limited, machine learning is of value in picking up a wider variety of ways of expressing symptoms. Of course, this depends on a) the coverage of the rules against which the SVM is being contrasted, and b) the confidence threshold of the SVM and other relevant parameters. However, this effect persisted even after varying the confidence threshold of the SVM quite substantially.

Optimizing precision presented more difficulties than improving recall. Varying the confidence threshold of the SVM to improve recall tended to cost more in recall than was gained in precision, so rule-based approaches were employed. However, it is much easier to specify what patterns do indicate a particular symptom than list all the ways in which the symptom might *not* be referred to. Symptoms varied a lot with respect to the extent of the precision problem. In particular, abstract thinking, which relies a lot on the word "concrete", which may appear in many contexts, posed problems, as did emotional withdrawal, which is often indicated by quite varied use of the word "withdrawn", which may occur in many contexts. Other symptoms, whilst easier than abstract thinking and social withdrawal, are also variable in the way they are expressed. Mood, for example, is often described in expressive and indirect ways, as is poverty of speech. On the other hand, mutism is usually very simply described, as is eye contact. It is an aid in this task that medical professionals often use quite formalized and predictable ways of referring to symptoms.

Aside from that, task difficulty depended to a large extent on the number of categories into which symptoms may be split. For example, the simple "mute" category is easier than eye contact, which may be good, intermediate or poor, with intermediate often being difficult to separate from good and poor. Likewise, speech may show poverty or be normal or abnormal, with many different types of problem indicating abnormality.

We chose to use an existing open-source language engineering toolkit for the creation of our applications; namely GATE (Cunningham et al., 2011). This approach enabled rapid prototyping, allowing us to make substantial progress on a large number of symptoms in a short space of time. The first version of a new symptom was added using default tool settings and with no additional programming. It was often added to the repertoire in under an hour, and although not giving the best results, this did achieve a fair degree of success, as seen in Table 1 which presents the machine learning-only results. In the case of the simpler symptoms (apathy and social withdrawal), this initial system gave sufficient performance to require no further development.

Additional training data was obtained for five symptoms, by presenting labelled sentences with low classifier confidence to the annotator (Table 2). Although this did improve performance, it is unclear whether this was due to an increase in training data alone, or whether concentrating on the low confidence examples made a difference. The annotator did, however, report that they found this approach easier, and that it took less time than annotating full documents for each symptom.

6 Conclusion

In conclusion, a good degree of success has been achieved in finding and classifying negative symptoms of schizophrenia in medical records, with precisions in the range of 0.8 to 0.99 being achieved whilst retaining recalls in excess of 0.5 and in some cases as high as 0.96. The work has unlocked key variables that were previously inaccessible within the unstructured free text of clinical records. The resulting output will now feed into epidemiological studies by the NIHR Biomedical Research Centre for Mental Health.

We asked whether off-the-shelf language engineering software could be used to build symptom extraction applications, with little or no additional configuration. We found that it is possible to create prototypes using such a tool, and that in the case of straightforward symptoms, these perform well. In the case of other symptoms, however, language engineering skills are required to enhance performance. The best results were obtained by adding hand-crafted rules that dealt with weakness in the machine learning.

References

N. C. Andreasen. 1983. *Scale for the Assessment of Negative Symptoms*. University of Iowa Press, Iowa City. Cited by 0000.

H. Cunningham, D. Maynard, and V. Tablan. 2000. JAPE: a Java Annotation Patterns Engine (Second Edition). Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield, Sheffield, UK, November.

H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M.A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters. 2011. *Text Processing with GATE (Version 6)*.

Hamish Cunningham, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. 2013. Getting more out of biomedical documents with gate's full lifecycle open source text analytics. *PLoS Computational Biology*, 9(2):e1002854, 02.

Andrea C Fernandes, Danielle Cloete, Matthew TM Broadbent, Richard D Hayes, Chin-Kuo Chang, Angus Roberts, Jason Tsang, Murat Soncul, Jennifer Liebscher, Richard G Jackson, Robert Stewart, and Felicity Callard. 2013. Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. *BMC Medical Informatics and Decision Making*. Accepted for publication.

T. Greenhalgh, H. W. Potts, G. Wong, P. Bark, and D. Swinglehurst. 2009. Tensions and paradoxes in electronic patient record research: a systematic literature review using the meta-narrative method. *Milbank Quarterly*, 87(4):729–788, Dec.

S R Kay, A Fiszbein, and L A Opler. 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia bulletin*, 13(2):261–276. Cited by 8221.

Y. Li and J. Shawe-Taylor. 2003. The SVM with Uneven Margins and Chinese Document Categorization. In *Proceedings of The 17th Pacific Asia Conference on Language, Information and Computation (PACLIC17)*, Singapore, Oct.

Y. Li, K. Bontcheva, and H. Cunningham. 2005. Using Uneven Margins SVM and Perceptron for Information Extraction. In *Proceedings of Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*.

Yaoyong Li, Kalina Bontcheva, and Hamish Cunningham. 2009. Adapting SVM for Data Sparseness and Imbalance: A Case Study on Information Extraction. *Natural Language Engineering*, 15(2):241–271.

Keith Lloyd, Matteo Cella, Michael Tanenblatt, and Anni Coden. 2009. Analysis of clinical uncertainties by health professionals and patients: an example from mental health. *BMC Medical Informatics and Decision Making*, 9(1):34.

S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, and J.F. Hurdle. 2008. Extracting information from textual documents in the electronic health record: A review of recent research. *IMIA Yearbook of Medical Informatics*, pages 128–144.

P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman. 2011. Natural language processing: an introduction. *J Am Med Inform Assoc*, 18(5):544–551.

S. M. Powsner, J. C. Wyatt, and P. Wright. 1998. Opportunities for and challenges of computerisation. *Lancet*, 352(9140):1617–1622, Nov.

Francisco S. Roque, Peter B. Jensen, Henriette Schmock, Marlene Dalgaard, Massimo Andreatta, Thomas Hansen, Karen Seby, Sren Bredkjr, Anders Juul, Thomas Werge, Lars J. Jensen, and Sren Brunak. 2011. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput Biol*, 7(8):e1002141, 08.

S. T. Rosenbloom, J. C. Denny, H. Xu, N. Lorenzi, W. W. Stead, and K. B. Johnson. 2011. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *J Am Med Inform Assoc*, 18(2):181–186.

Donia Scott, Rossano Barone, and Rob Koeling. 2012. Corpus annotation as a scientific task. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Burr Settles. 2012. *Active Learning*. Morgan and Claypool.

Sunghwan Sohn, Jean-Pierre A Kocher, Christopher G Chute, and Guergana K Savova. 2011. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *Journal of the American Medical Informatics Association*, 18(Suppl 1):i144–i149.

Robert Stewart, Mishael Soremekun, Gayan Perera, Matthew Broadbent, Felicity Callard, Mike Denis, Matthew Hotopf, Graham Thornicroft, and Simon Lovestone. 2009. The South London and Maudsley NHS foundation trust biomedical research centre (SLAM BRC) case register: development and descriptive data. *BMC Psychiatry*, 9:51–62.

4.2 TextHunter

The successes of the initial negative symptom work prompted methodological discussions about how the classification performance of SVMs might be leveraged to tackle the bottlenecks of rules based approaches. Here, we theorised that ML might circumvent the need for expensive knowledge engineer time, on the assumption that the classification performance of a specific concept extraction model was positively correlated with the volume of training data associated with it. This presented several operational advantages. First, training data could be created by CRIS project owners without extensive knowledge engineering training, who also had vested interests in the success of a given IE application. Second, CRIS project owners are likely to be domain experts for the specific concept they wish to extract, allowing them to directly define concept specific annotation guidelines without the need for extensive cross-domain communication. Finally, by standardising the inputs and outputs of an IE system, the costs of day-to-day operational management of research dataset construction would be reduced.

However, many issues remained to be resolved, such as model optimisation, active learning implementation and defining the domain model. As a result of this work, a range of enhancements to the original methodology was produced. This in turn led to the development of the TextHunter software suite. In November 2014, a peer reviewed publication describing the software and methodology was presented at the AMIA 2014 conference [3]:

Author contributions:

- Richard Jackson: Wrote all code, wrote the paper, developed the methodology, assisted with annotation, completed the analysis
- Michael Ball: Assisted with annotation
- Rashmi Patel: Assisted with annotation and provided user experience feedback
- Richard Hayes: Assisted with annotation and provided user experience feedback
- Richard Dobson: Provided supervisor support
- Robert Stewart: Provided supervisor support and assisted with annotation

TextHunter – A User Friendly Tool for Extracting Generic Concepts from Free Text in Clinical Research

Richard G. Jackson MSc[1], Michael Ball MSc[1], Rashmi Patel BMBCh[1], Richard D. Hayes PhD[1], Richard J.B. Dobson PhD[1], Robert Stewart MD[1]

[1]King's College London (Institute of Psychiatry), London, UK

Abstract

Observational research using data from electronic health records (EHR) is a rapidly growing area, which promises both increased sample size and data richness - therefore unprecedented study power. However, in many medical domains, large amounts of potentially valuable data are contained within the free text clinical narrative. Manually reviewing free text to obtain desired information is an inefficient use of researcher time and skill. Previous work has demonstrated the feasibility of applying Natural Language Processing (NLP) to extract information. However, in real world research environments, the demand for NLP skills outweighs supply, creating a bottleneck in the secondary exploitation of the EHR. To address this, we present TextHunter, a tool for the creation of training data, construction of concept extraction machine learning models and their application to documents. Using confidence thresholds to ensure high precision (>90%), we achieved recall measurements as high as 99% in real world use cases.

Introduction

The increasing use of electronic health records (EHR) provides potentially transformative opportunities for clinical research in the breadth and depth of data contained within them. However, unstructured clinical notes are often the most valuable source of phenotypic/contextual information because of limitations in the scope and acceptability of structured fields. In response to this challenge, Natural Language Processing (NLP) has been employed to extract appropriate assertions in a structured format amenable to the needs of researchers[1]. While significant success has been achieved in many areas, the demand for ever more variables to be extracted from the clinical narrative is currently bottlenecked by the limited supply of technical skills[2]. For example, rule based approaches to novel problems are often effective, but require a certain degree of technical knowledge and experience, which can be too time-consuming and thus expensive to produce in high volume. Proposed solutions include ontological or dictionary mapping techniques, which are appropriate where there are well-constructed resources; however, the standards imposed by these may not be easy to adapt to real world clinical sub-languages[3], or where there is controversy about the appropriate use of clinical language[4]. Machine Learning (ML) approaches are an increasingly popular means of circumventing rule based systems, but require even more technical expertise and are limited by the availability and ease of creating appropriate training data[5,6]. Finally, although progress has been made in the development of publicly available corpora for evaluating different clinical NLP methodologies[7], these offer no guarantee that the performance obtained by models trained on such data will provide a generalizable solution (for example, for work on EHRs in different medical domains, dialects, languages, or work cultures)[8].

These issues form barriers to progress for groups who have access to unstructured clinical data, but do not have sufficient technical capabilities to trial the wealth of information extraction techniques on offer. In recent years this has prompted the development of tools such as Arc[9] to democratize access to generic information extraction capabilities. However, there are currently no free tools available that offer a full end-to-end solution for concept level extraction, including the principle tasks of:

1) Extracting instances of concepts from a database or large collection of documents

2) Creating sufficient training data specific to a concept to enable a machine learning approach

3) The configuration and testing of an (ML) algorithm for the given concept

4) The application of the model to the entire document set of interest, and the subsequent export of results into a familiar format

In order to make concept extraction technologies accessible to groups without informatics support, we have developed the TextHunter tool to address these tasks.

Methods

Data: The South London and Maudsley mental health case register

The South London and Maudsley NHS Trust (SLAM) is the largest mental health organization in Europe, and is a virtual monopoly provider of mental health services to 1.2 million individuals within its geographical catchment area (Lambeth, Southwark, Lewisham and Croydon boroughs in South London). In 2007-08, funding from the British National Institute for Health Research supported the development of the Clinical Record Interactive Search (CRIS) database. CRIS operates as a pseudonymized version of SLAM's EHR system, accessible for researchers via its distinctive, patient-led information governance model[10]. CRIS houses more than 230,000 de-identified patient records, which in turn represent over 20 million free text documents. The CRIS system continues to grow at a rate of approximately 170,000 free text documents per month. Clinical information documented in unstructured text is of particular value in mental health research where there is an increasing emphasis on using dimensional symptom scales to define mental illness rather than discrete diagnostic categories[11–14]. While CRIS also has large amounts of data contained within structured fields, the development of TextHunter was precipitated by the needs of many disparate groups of researchers who require access to the wealth of additional information contained within the clinical narrative.

TextHunter System Description

TextHunter is a program that guides a user through all of the required processes to create and apply a concept extraction model for a selection of documents from start to finish. It performs six important tasks, the end result of which delivers a structured representation of a concept. Its intended use case is typically phenotype cohort identification, although it can be employed for more generic purposes. The program is built from open source libraries, and uses the GATE library as its core NLP engine[15]. The ML element uses the Support Vector Machine (SVM) based 'Batch Learning' plugin supplied with GATE[16]. In consideration of the rigorous information governance requirements of clinical data, TextHunter is designed to operate as a standalone 'offline' program on desktop hardware, although its multithreaded design enables its deployment on more powerful workstations/virtual machines to handle larger datasets. It is capable of connecting to commercial database environments such Microsoft SQL Server to process massive datasets, but for succinctness, only its standalone operation mode is described here.

The underlying principle of TextHunter is 'Find, Annotate, Build, Apply' - respectively addressing the four key problems described above. The integration of these concepts into a single system creates the possibility of providing lay users with access to more advanced ML techniques, such as active learning. Each phase of the TextHunter pipeline is described below:

1. Search Phase

This phase addresses task 1). The first stage of the TextHunter pipeline requires a user to define a list of keywords, regular expressions and/or phrases to describe their concept of interest. The user then directs the program to a directory holding the text files of interest. Upon executing the 'search' phase, each document is scanned for mentions of the user's expressions. When a mention is identified, a short section of text consisting of multiple sentences, including the sentence where the concept mention was found, and up to two sentences either side of the sentence of interest is extracted. This is stored in an embedded file based database, along with a copy of the underlying document. Deconstructing documents in this way facilitates the downstream management of text instances for annotation and classification.

2. Annotation Phase

This phase addresses task 2). The user is directed to TextHunter's annotation interface, which has been specifically designed for the rapid annotation of concept instances. We define an instance as a group of one to five sentences centered on a concept keyword, and its classification as defined below:

i) Positive – the example is a relevant hit and is an appropriate positive example of the user's concept

ii) Negative - the example is a relevant hit and is an appropriate negated example of the user's concept

iii) Unknown - the example is a relevant hit but the user is unable to ascertain the correct classification, or the example is irrelevant

In this phase, the user is required to produce a 'test' corpus for model validation (typically of 100-300 instances), which are randomly selected from all instances in the document set. This is followed by the production of a 'seed' corpus to be used in training models. This also numbers about 100-300 instances, but is enriched by ensuring no identical instances are present. In real world clinical datasets, the required semantic context that enables the classification of a concept instance may cross sentence boundaries. To ensure appropriate features are available for training, the user can specify the required 'context' (up to two sentences before and two sentences after) needed to make the classification, centered on the sentence containing the concept keyword. These boundaries are arbitrarily chosen by the GATE sentence splitter module, although we expect that only in very rare cases will more than five sentences be required to express medical concepts as they are normally found in EHRs.

3. Feature selection/Model Building Phase

This phase addresses task 3). Here, TextHunter builds and evaluates a range of models against the task, using different features and SVM parameters each time. The default feature vector used by TextHunter is a classic bag of words using part-of-speech tags and token stems from the user specified context around a concept. When applying a model to unseen data, TextHunter creates feature vectors from up to six different combinations of sentences around the sentence containing the concept term. The classification resulting from the feature vector producing the highest overall confidence is chosen as the result. In addition, TextHunter has a modular design that allows developments from the clinical NLP community to be integrated into its core pipeline via GATE creole plugins. Currently, TextHunter takes features of the GATE implementation of the ConText algorithm[17], which uses hand crafted rules to determine whether a concept is negated, temporally irrelevant or refers to a subject other than the patient. Stop word removal is also explored during feature selection.

Cross validation of the training data is used to mitigate the dangers of overfitting the model to a small amount of data. The model producing the best F1 score is taken forward for testing against the human labeled 'test' corpus, which is never used in model training. A range of easy to interpret output files are produced, containing estimates of 'real world' performance the user might expect.

4. Application Phase

This phase addresses task 4). This phase allows the user to apply the best performing model to all instances of text in their dataset, as captured in the search phase. As with the model building phase, combinations of sentences are tested around the concept. The classification that results from the combination with the highest confidence is chosen as the final result. Once this stage is complete, the user may export the output into several formats.

5. Active Learning Phase (optional)

Conceptually, active learning is an iterative process whereby an ML algorithm selects instances that it has difficulty classifying and presents them to a human annotator for labeling. These are then fed back into the model, with the intention that the new model arising will be better at classifying similar, difficult examples. TextHunter supports a 'simple margin' inspired method of active learning[18]. A seed model is constructed from randomly selected instances of text, as described above. This model is then applied to a large sample of the entire population of relevant text instances. For each classification the model makes, it also assigns a level of certainty, between -1 and +1. Theoretically, highly positive scores are representative of easy to classify 'positive' instances, whereas highly negative scores are representative of easy to classify 'negative' or 'unknown' instances. Instances with a certainty score close to 0 are thus 'difficult', and presented to the user for labeling in order to retrain the classifier.

Use cases

To evaluate the performance of TextHunter, we defined three real world use cases of concept extraction. Examples of search expressions and typical instances for each use case are detailed in Table 1:

Case Study 1: Cannabis Smoking

Cannabis use has been indicated as a potentially aggravating factor in patients suffering from mental illness[19]. Through the vast amount of electronic documentation generated in the course of patient care, we attempted to identify a patient's cannabis smoking status based upon reports by mental health professionals. The CRIS database contains intra-profession clinical correspondence style documents and clinical notes resulting from patient contact. Each type of document may contain references to cannabis usage by the patient. In this study, our objective was to use TextHunter to build a classifier to identify current or historical cannabis usage. We conducted a review of the most common nouns and slang terms used to describe cannabis in SLAM, to produce a list of expressions which formed the basis for finding instances to classify. A psychiatrist then produced multiple sets of annotations using the standard TextHunter procedure, making use of the active learning functionality. Although it was not possible to double annotate the training data, we adopted a restrictive manual coding strategy in order to allow as little subjectivity as possible (for example, by classifying mentions pertaining to future events, or tangential/circumstantial references into our predefined 'unknown' class).

Case Study 2: Psychosis Symptomatology

Patients suffering from psychosis can exhibit a wide range of symptoms, which in turn inform the nature of their treatment plan. Common tools to quantify symptomatology in psychosis include such instruments as the Positive and Negative Symptom Scale and the Clinical Assessment Interview for Negative Symptoms[12,13]. These depend on an assessment of the patient's presentation in regard to a wide range of possible symptoms. Our previous work to capture some of these from clinical notes with ML approaches has been described[20,21]. In this case study, we used TextHunter to capture two additional symptoms: delusional symptoms and evidence of hallucinations, using the standard TextHunter workflow. The annotated data for 'delusions' were generated by a clinical informatician, with a random sample checked for accuracy and consistency by a psychiatrist. In the case of 'hallucinations', all annotations were generated by a public health physician. In both cases, the restrictive coding strategy as described above was employed.

Table 1: Search expressions and examples of instances. Theoretical patient identifiers masked by ZZZZZ.

Case Study	**Examples of subword patterns (case insensitive) for search phase**	**Fictitious examples of instances (Parentheses indicates typical labeling by human annotator)**
Cannabis smoking	cannab hash weed pot	ZZZZZ told me that he continues to smoke cannabis only no other illicit drugs. **(positive)** ZZZZZ has no history of amphetamine or cannabis use. **(negative)**
Psychosis symptomatology	delusio hallucina	She is continuing to experience hallucinations and is becoming increasingly distressed by these. **(positive)** Staff observed him to rambling and delusional, repeating himself and his gait was abnormal and more pronounced. **(positive)**

Case Study 3: Ethnicity

Ethnicity is a key variable in many epidemiological and clinical studies. Although ethnicity can theoretically be captured via the structured elements in SLAM's EHR system, in reality, it is often not recorded in the course of routine clinical practice. However, as with many other variables, ethnicity is often referenced in clinical free text. The purpose of this case study was therefore to classify instances of text describing a patient's ethnicity, into one of 17 ethnic groups. A range of terms was selected in association with each ethnic group, and a 'positive' classification was made if the context for the term was suggestive of the patient belonging to that group. A single researcher produced the annotated dataset for training/testing, using a similarly restrictive coding strategy.

In each case study, a sample of the evaluation instances were double annotated by an individual in a related profession to generate inter-annotator agreement statistics.

Results

In all case studies, we used 10 fold cross validation for the model building phase, which took approximately one hour on a desktop computer with a Core 2 Duo E7500 processor.

In the cannabis smoking study, we used 13 terms to capture cannabis mentions. The CRIS database yielded 663,979 mentions of cannabis. For the psychosis symptomatology study, the search phase found 603,818 mentions of delusions, and 703,996 mentions of hallucinations. Each symptom was represented by a single term in the search phase. Finally, there were 3,444,435 mentions of concepts potentially related to ethnicity, resulting from 277 terms commonly used to define our 17 ethnic identities.

Traditionally, the performances of information extraction algorithms in NLP are described in terms of precision, recall and the F1 statistic. However, the high level of noise commonly associated with EHR based observational research necessitates the capture of high quality data in order to generate clearly defined cohorts. This data quality requirement restricts the use of automated concept extraction techniques to those that can be shown to have a high true positive rate, relative to the inherent predictive value of a mention of a concept. For example, a mention of a cannabis synonym will refer to a patient's current or past use 70% of the time, whereas a mention of a term denoting ethnicity will refer to a patient's actual ethnicity only 20% of the time (Table 1). A further consideration of the real world viability of a given model is the longitudinal nature of the electronic health record. A patient may have numerous contacts with a health service over a number of years, creating multiple instances of time independent concepts. For example, a patient may have multiple references to their cannabis consumption habits, especially if it is identified as a factor in their illness. Similarly, a patient's ethnicity may be described in service referral letters generated during the course of their care. Only one positive instance needs to be captured precisely for a high quality output to be achieved. However, spurious data points are more problematic. Given these factors, it is more practical to develop information extraction tools that favor precision over recall in most use cases. For this reason, in Table 2 we describe the recall statistic at two arbitrarily defined levels of precision (90% and 95%), which are identified by filtering the classified instances in the test set via the classification confidence threshold. We present Receiver-Operator Characteristic (ROC) plots for each case study in Figure 1. For brevity, we only report the highest F1 achieved without any confidence filtering (note, this is not necessarily the same model that achieves the highest recall at the 90%/95% precision threshold).

The best performance was seen in the hallucinations case study, with over 97 % recall obtained at the 95% precision threshold. The worst performance was observed in the ethnicity study, where recall reached only 9% at 90% precision, and declined with further training.

Different problems required different features in order to obtain the best overall result. In Table 3, we present the types of features that were found to be most useful in each case study.

The rate of training data production varied moderately between the studies, the slowest recorded at approximately 100 instances labeled per hour, and the fastest at roughly 230 instances per hour. Since different individuals annotated each study, further comparisons were not possible. Anecdotal reports from the annotators suggested that the process of annotating instances selected via active learning was slower than the randomly selected instances in the seed set.

Table 2: Performance statistics for TextHunter 'positive' instances ('unknown' and 'negative' instances are grouped together). [1]Observed Agreement and Cohen's Kappa. [2]Baseline precision assumes presence of keyword is a 'positive' instance (by definition, recall is 100%), and provides a measure of how predictive a mention is of a concept without any processing applied. P = precision, R = recall, F1 = harmonic mean of precision and recall. [3]Parentheses indicate count of training instances in the model building phase (subsequent active learning iterations increase the number of training instances available). [4]Recall measured at precision levels of 90% and 95%, attained by confidence filtering.

Case Study	**Inter annotator agreement** [1,3]	**Test Instances**	**Baseline precision**[2]	**Seed data base performance**[3]	**Seed data Recall**[3,4]		**Active learning iteration 1 recall**[3,4]		**Active learning iteration 2 recall**[3,4]		**Approximate total annotator time spent creating training data**
					90	**95**	**90**	**95**	**90**	**95**	
Cannabis smoking	88% 0.76 (211)	233	75%	P = 81% R = 95% F1 = 0.87 (478)	45	38	68	52	72	53	~10 hours
					(478)		(1 329)		(1 835)		
Delusions	95% 0.91 (110)	206	68%	P = 89%/ R = 99%/ F1 = 0.93 (708)	95	87	N/A		N/A		~4 hours
					(708)						
Hallucinations	89% 0.78 (117)	131	70%	P = 93% R = 99% F1 = 0.96 (150)	99	97	99	97	N/A		~7 hours
					(150)		(914)				
Ethnicity	97% 0.94 (201)	650	20%	P = 82% R = 75% F1 = 0.78 (396)	9	9	3	3	N/A		~3 hours
					(396)		(805)				

Table 3: Additional features used in best performing model delivering >90% precision

Case Study	Best Model ID	ConText used?	Stop words removed?	SVM Cost	SVM kernel type
Cannabis Smoking	128	No	No	0.6	polynomial
Delusions	136	No	No	0.6	polynomial
Hallucinations	88	Yes	No	0.5	polynomial
Ethnicity	24	Yes	Yes	0.7	polynomial

Discussion:

In our analysis, we used TextHunter to extract a diverse set of concepts that are typically in demand in clinical research environments. We arbitrarily set two desired precision standards, and adopted strategies to try to maximize the recall given this requirement. Three of the four test cases reached over 70% recall at the lower precision cut-off of 90%. We do not attempt to tackle the question of what constitutes acceptable performance for research applications here. Nevertheless, we have confidence that the range of case studies investigated here establishes a

proof of concept in enabling end users to create and deliver information extraction solutions independently of significant NLP expertise.

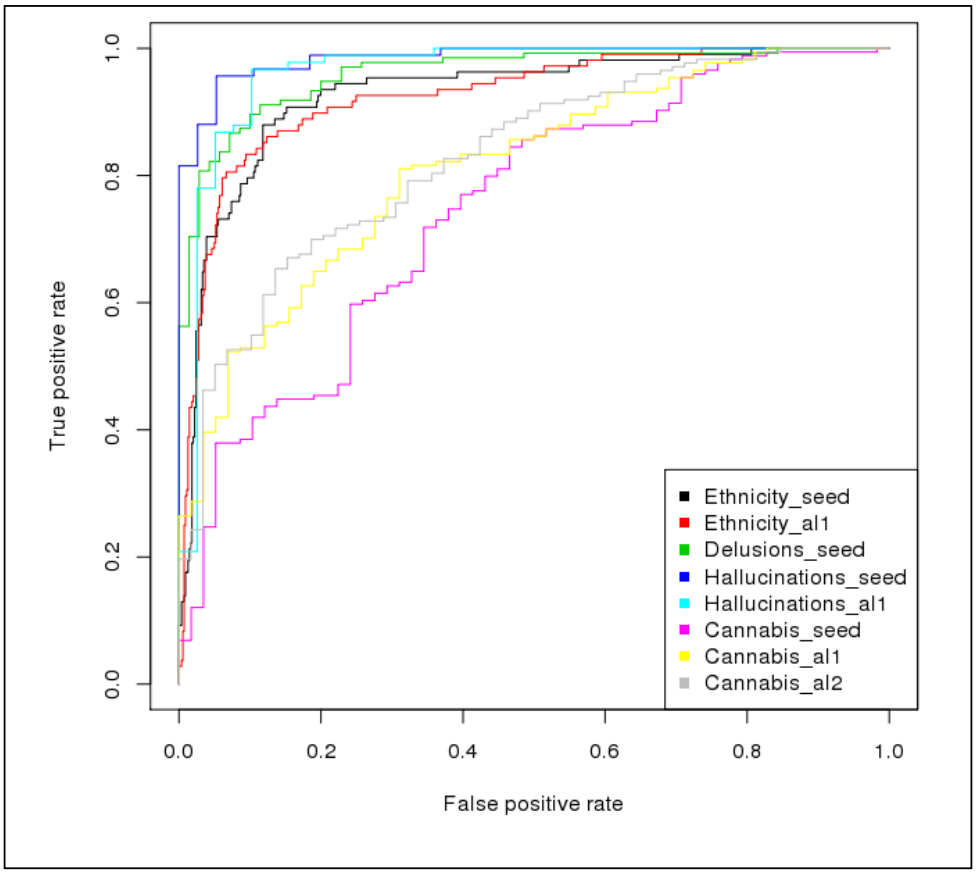


Figure 1: Receiver Operator Characteristic for TextHunter models on 'test' data, generated with SVM confidence thresholds.

Given our limited range of test cases, the SVM parameters and additional features used varied greatly, even between the two conceptually similar problems explored in psychosis symptomatology. This substantiates our approach of testing a range of models to find the best solution for a given problem. However, a predominant factor in the algorithms' ability to reach higher levels of recall is the predictive value that a simple mention of a concept produces (i.e. how likely a human annotator is to label a randomly selected mention of a concept as 'positive'). For instance, the ROC curve produced for the ethnicity study compares favorably with that of the cannabis study, and we achieved a substantial performance benefit over the baseline precision for our list of ethnicity terms. However, because of our self-imposed requirement of a minimum 90% precision, the recall for ethnicity falls very quickly as this threshold is approached. Intuitively, in high noise datasets where 'positive' mentions of a concept are rarer, the concept extraction problem is significantly more challenging. In addition, the low predictive value of ethnicity terms means the 'positive' class will be less represented than the 'negative' or 'unknown' classes in the model. Currently, TextHunter makes no adjustment for unbalanced classes, and future work could investigate mitigation strategies for this, such as using uneven margins [22]. It should also be noted that we were required to use many more terms to capture mentions of ethnicity, which may be indicative of the inherent difficulty of defining concepts that are largely social constructs.

In the case of the cannabis study, we were able to improve the model substantially by providing additional training data through active learning. We did not try to quantify the added benefit of adopting an active learning

methodology over randomly selecting new instances. However, others have previously demonstrated that active learning can accelerate the development of machine learning models in clinical NLP[18,23,24]. Active learning did not produce an additional benefit in the hallucinations case study, although the model resulting from the seed data had already produced a very high F1 statistic. Here, our application of confidence filters was not required, as the performance of the model generated from the seed annotations surpassed our precision requirement of 95%. In the case of ethnicity, adopting an active learning approach noticeably depreciated the quality of the model. To investigate, we conducted a subjective review of the instances that active learning retrieved. This revealed that many were incoherent strings of text, seemingly resulting from jumbled emails, faxes and other malformed documents. Since these were not representative of natural language, their inclusion in training the model possibly introduced more noise than benefit. Previous reports have highlighted the difficulties of applying general NLP tools on clinical text[8,25], and we suspect that this scenario is not uncommon in real world EHR systems. One possible mitigation strategy would be to employ document classification methods to filter out malformed documents and/or a more sophisticated active learning methodology, such that new training data are more representative of the instances of interest. Nevertheless, an SVM approach as implemented in TextHunter appears to be valid for simple concepts that tend be succinctly expressed - for example, if it can be defined with a relatively short list of keywords, is not over-complicated by frequent ungrammatical usage (such as in lists or questionnaire text) and has a baseline precision of at least 60%.

It was not practical to double annotate our training data fully, so we are only able to provide inter-annotator agreement (IAA) statistics for a subset of the total test set in each case study. Despite our limited set, our data suggest relatively high levels of agreement, highlighting a high degree of objectivity in the expression of concepts in clinical text. However, clinical constructs in mental illness are often subtle. Initial reports from annotators in each case study suggested that the annotation process itself influenced their own views on the interpretation of notes created by others. Specifically, the exposure to a wide range of writing styles from other clinicians may introduce unforeseeable subjectivity into the annotation process. Regardless, methods that place subject matter experts (rather than NLP specialists) in the role of defining a concept are likely to be less subjective, as any subjectivity introduced by the annotation process will likely be compounded by attempting to convey the subtleties to a non-expert third party. Any clinical subjectivity may then be mitigated by a process of iterative discussion and re-annotation to produce well defined annotation guidelines. A potentially useful future development of TextHunter may be to incorporate a model of clinical data, such as the Clinical Element Model[26]. This would encourage the re-use of standard definitions of concepts, thus promoting greater interoperability with NLP tools.

A notable shortcoming of the TextHunter methodology was the ethnicity case study, which had the highest Kappa statistic but the lowest F1 score from the seed data. This highlights the divide between human and machine interpretation, and the need for more complex reasoning systems to resolve more difficult problems.

Conclusion

The requirement to develop this software was driven by an imbalance between the demand for concept extraction and the supply of skilled individuals capable of delivering solutions to the needs of researchers. We have shown that it is feasible to package an appropriate suite of tools into a simple interface, and that this enables researchers to produce concept extraction models without input from NLP specialists. TextHunter uses a flexible SVM based algorithm as a generic, user friendly information extraction capability. We have validated the methodology with a variety of typical problems, and produced high precision and relatively high recall models. Although it is not suitable for all tasks, we argue that the ‘solve small problems quickly’ approach to information extraction is appropriate for many types of variable likely to be of interest to researchers, and offers the attractive advantage of rapidly generating models that have been trained on data sourced from the intended target. Finally, the simple annotation interface enables a rapid annotation process, with labeled data stored in a standard, reusable format. The pipeline style operation of GATE and the open source licence of TextHunter should encourage the future development of additional features to improve performance and expedite its use on more complex NLP problems.

TextHunter is available at https://github.com/RichJackson/TextHunter

Funding/Support Acknowledgement

RJ, RD and RS are part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. MB is supported by the BRC Nucleus jointly funded by the Guy's and St Thomas' Trustees and the South London and Maudsley Trustees. RP is supported by a Medical Research Council Clinical Research Training Fellowship. RH is funded by a Medical Research Council (MRC) Population Health Scientist Fellowship. RD and RS are joint last authors on this work.

References

1. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. Journal of the American Medical Informatics Association. 2013 Nov 7;21(2):221–30.

2. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. Journal of the American Medical Informatics Association. 2011 Aug 16;18(5):540–3.

3. Demner-Fushman D, Mork JG, Shooshan SE, Aronson AR. UMLS content views appropriate for NLP processing of the biomedical literature vs. clinical text. Journal of Biomedical Informatics. 2010 Aug;43(4):587–94.

4. Chmielewski M, Bagby RM, Markon K, Ring AJ, Ryder AG. Openness to Experience, Intellect, Schizotypal Personality Disorder, and Psychoticism: Resolving the Controversy. J Pers Disord. 2014 Feb 10;

5. Afzal Z, Schuemie MJ, van Blijderveen JC, Sen EF, Sturkenboom MCJM, Kors JA. Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records. BMC Med Inform Decis Mak. 2013;13:30.

6. Khor R, Yip W-K, Bressel M, Rose W, Duchesne G, Foroudi F. Practical implementation of an existing smoking detection pipeline and reduced support vector machine training corpus requirements. Journal of the American Medical Informatics Association. 2013 Aug 6;21(1):27–30.

7. Zhai H, Lingren T, Deleger L, Li Q, Kaiser M, Stoutenborough L, et al. Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. J Med Internet Res. 2013;15(4):e73.

8. Patterson O, Hurdle J. Document Clustering of Clinical Narratives: a Systematic Study of Clinical Sublanguages. AMIA Annu Symp Proc. 2011.

9. D'Avolio LW, Nguyen TM, Goryachev S, Fiore LD. Automated concept-level information extraction to reduce the need for custom software and rules development. Journal of the American Medical Informatics Association. 2011 Jun 22;18(5):607–13.

10. Stewart R, Soremekun M, Perera G, Broadbent M, Callard F, Denis M, et al. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. BMC Psychiatry. 2009;9:51.

11. Adam D. Mental health: On the spectrum. Nature. 2013 Apr 24;496(7446):416–8.

12. Kring AM. The Clinical Assessment Interview for Negative Symptoms (CAINS): Final Development and Validation. American Journal of Psychiatry. 2013 Feb 1;170(2):165.

13. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull. 1987;13(2):261–76.

14. Axelrod BN, Goldman RS, Alphs LD. Validation of the 16-item Negative Symptom Assessment. J Psychiatr Res. 1993 Sep;27(3):253–8.

15. Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. Prlic A, editor. PLoS Computational Biology. 2013 Feb 7;9(2):e1002854.

16. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011 Apr 1;2(3):1–27.

17. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of Biomedical Informatics. 2009 Oct;42(5):839–51.

18. Koller D, Tong S. Support vector machine active learning with applications to text classification. The Journal of Machine Learning Research. 2001;2:45–66.

19. Moore TH, Zammit S, Lingford-Hughes A, Barnes TR, Jones PB, Burke M, et al. Cannabis use and risk of psychotic or affective mental health outcomes: a systematic review. The Lancet. 2007 Jul;370(9584):319–28.

20. Gorrell G, Jackson R, Roberts A. Finding Negative Symptoms of Schizophrenia in Patient Records. Proc NLP Med Biol Work (NLPMedBio). Hissar, Bulgaria; 2013. p. 9–17.

21. Patel R, Jayatilleke N, Jackson R, Stewart R, McGuire P. Investigation of negative symptoms in schizophrenia with a machine learning text-mining approach. The Lancet. 2014 Feb;383:S16.

22. Li Y, Bontcheva K, Cunningham H. Adapting SVM for data sparseness and imbalance: a case study in information extraction. Natural Language Engineering. 2008 Dec 18;15(02):241.

23. Chen Y, Carroll RJ, Hinz ERM, Shah A, Eyler AE, Denny JC, et al. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. Journal of the American Medical Informatics Association. 2013 Jul 13;20(e2):e253–e259.

24. Figueroa RL, Zeng-Treitler Q, Ngo LH, Goryachev S, Wiechmann EP. Active learning for clinical text classification: is it better than random sampling? Journal of the American Medical Informatics Association. 2012 Jun 15;19(5):809–16.

25. Barrett N, Weber-Jahnke JH. Applying natural language processing toolkits to electronic health records - an experience report. Stud Health Technol Inform. 2009;143:441–6.

26. Wu ST, Kaggal VC, Dligach D, Masanz JJ, Chen P, Becker L, et al. A common type system for clinical natural language processing. Journal of Biomedical Semantics. 2013;4(1):1.

4.3 Supplemental System Description

Since the original TextHunter manuscript was published, work has focused on additional improvements to this methodology, and their implementation into the TextHunter software. The rest of this chapter describes the current status of the TextHunter Software.

TextHunter is built around the GATE framework [95]. In addition to offering a Graphical User Interface (GUI), GATE features Application Programming Interfaces (APIs) that enable its extension and integration into other NLP software applications.

The TextHunter methodology consists of a three stage process.

4.3.1 Step 1 - Information Retrieval

TextHunter formulates the IE problem as a sentence classification problem. Here, a given sentence containing a named entity of interest may belong to one of two classes: relevant or irrelevant. For instance, if we assume the entities of interest are 'hallucinations' and 'delusions', we would seek to classify all sentences containing such a named entity in a corpus of text. The first step of a TextHunter project is to identify the location of all sentences in the corpus which may be relevant to the desired concept. This may be thought of as an NER and/or IR problem. As we assume the user already has (or is capable of obtaining) sufficient domain awareness for a sufficiently broad range of terms to describe how their entity is expressed, we forego resource heavy NLP tasks such as UMLS dictionary lookup and lexical normalisation. Instead, it is sufficient to use basic string matching/regular expression techniques to find all instances (figure 4.1.

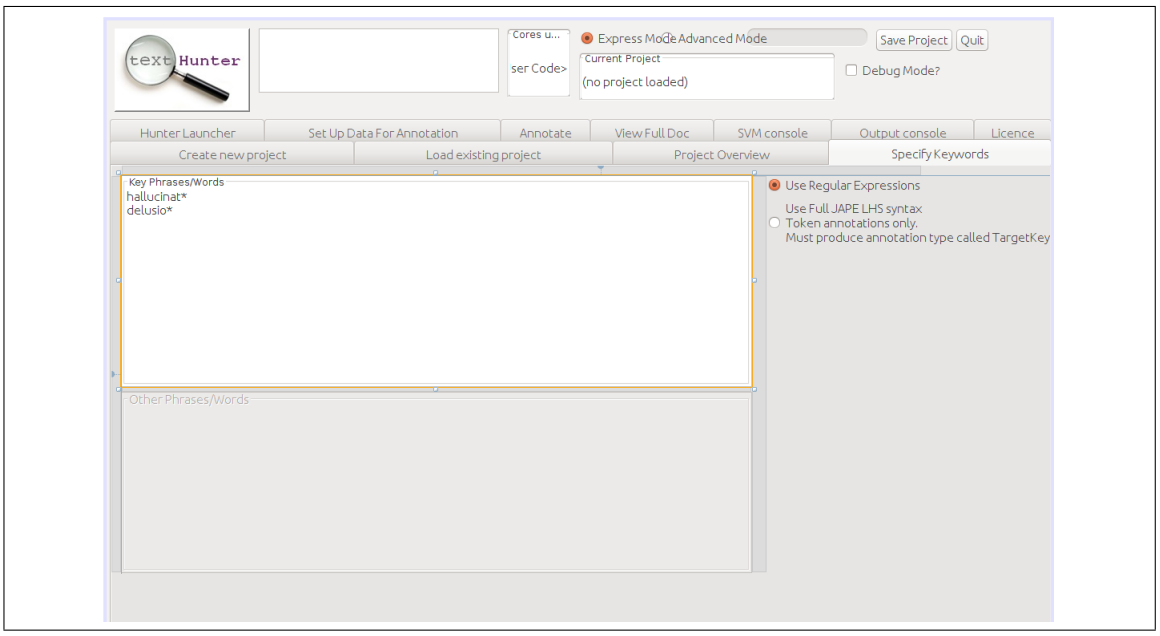


Figure 4.1: NER using simple regular expressions.

These keywords specify the search criteria. Once the appropriate keywords have been

entered, TextHunter connects to a datasource, which is either a file system directory or a database of documents. These documents are then tokenised, and regular expression matches on token strings are identified. For multi-word entities, sequential regular expressions or full JAPE left hand side syntax can be used[1]. Hits in the document repository are then persisted to a database, recording the document offsets (between character document locations) of the context of the hit. Formally, the data model is given in listing 4.1.

Listing 4.1: Data Model of TextHunter Projects in SQL Server Dialect

```
CREATE TABLE targetTableName (
        --document metadata identifying schema location of original document
         [id] [int] IDENTITY(1,1) NOT NULL PRIMARY KEY,
         [BrcId] [int] NOT NULL,
         [CN_Doc_ID] [varchar](30) NOT NULL,
         [src_table] [varchar](30) NOT NULL,
         [src_col] [varchar](30) NOT NULL,

        --instance metadata
         [match] [varchar](max) NULL,
         [numWords] [varchar](max) NULL,
         [annotStart] [varchar](max) NULL,
         [annotEnd] [varchar](max) NULL,
         [contextStart] [varchar](max) NULL,
         [contextEnd] [varchar](max) NULL,
         [contextString] [varchar](max) NULL,
         [GOLDSTANDARD] [varchar](max) NULL,
         [UPDATETIME] [varchar](max) NULL,

        --fields to hold human annotator classifications
         [keyObservation1] [varchar](max) NULL,
         [KEYPRIORITY1] [varchar](max) NULL,
         [keyObservation2] [varchar](max) NULL,
         [KEYPRIORITY2] [varchar](max) NULL,
         [comments] [varchar](max) NULL,



        --TextHunter pipeline classifications
         [mlObservation1] [varchar](max) NULL,
         [mlObservation2] [varchar](max) NULL,
```

[1]See GATE documentation for details www.gate.ac.uk

```
 [MLpriority] [varchar](max) NULL,
 [prob] [varchar](max) NULL,

-- ConText (Harkema et al) Variables
 [Experiencer] [varchar](max) NULL,
 [Temporality] [varchar](max) NULL,
 [Directionality] [varchar](max) NULL,
-- additional stuff
 [modifiers] [varchar](max) NULL,

 -- project name
 [project] [varchar](max) NULL
);
```

In a clinical context, applications that interact with databases in such a way confer a number of benefits. The administration of such databases are generally under the management of central hospital I.T. departments, meaning that they are likely to be subject to existing backup/disaster recovery and information governance processes. Since creating an NLP application may involve many hours of annotating data, ensuring the resulting data is secured in an appropriate way is favourable for business continuity.

A feature of the TextHunter system is the manner in which documents are decomposed into contexts once a hit is found. A context is defined as the NE tokens within the sentence (the target sentence), two sentences before and two sentences after. Decomposing a corpus of documents into a series of classifiable contexts provides a range of advantages in standardising inputs, making downstream processes more efficient and predictable. This helps to avoid out of memory errors potentially destabilising the system when erratic input documents are encountered (as is frequently the case in any large corpus that hasn't undergone some form of data cleaning). The disadvantage of this approach is that, at training and classification time, the full document is not available, potentially removing important information. However, as the TextHunter system is only designed to attempt relatively simple information extraction tasks - any construct that is modified by many complex co-references throughout a large document will likely require a highly customised NLP application.

In some clinical genres, the concept of a sentence is not well defined. For instance, free text fields within an EHR may require responses to specific questions, such as, "List the presenting symtomatology of the patient below". Here, one might expect responses to be highly telegraphic (see 2.2.7). Similarly telegraphic text might be expected in GP typed

notes. With few linguistic features to work with, the interpretation of TextHunter results is dependent upon the document context in which the results occur. In such cases, it would be important to have an awareness of the classification performance of the system upon different document types and if necessary, train additional models per document type.

As I define the IE task as a sentence classification problem, the sentence splitter used in the pipeline has special significance. A special configuration of the default GATE sentence splitter is used, that ignores the default behaviour of new line characters acting as sentence delimiters, as this was found to produce the most desirable sentence splits.

4.3.2 Step 2 - Annotation

Once the table is populated, it is possible to start annotating them according to guidelines determined by the user. This functionality is provided via a custom annotation interface, which retrieves annotations at random from the underlying database (figure 4.2).

The annotator is presented with only the relevant sentences as defined by their search criteria. Annotations consist of simply marking each context as 'relevant' or 'irrelevant. Annotators can also highlight between 0-4 of the surrounding sentences if they contain information pertinent to the classification decision. Annotations can be completed by using keyboard shortcuts, greatly increasing the speed at which data can be labelled.

Using this annotation method, two sub-corpora are created; a gold standard which is held out for final evaluation; and a training set, used for model optimisation (see below).

4.3.3 Step 3.i Model Building

Training data produced in Step 2 are then used to build the SVM classifier. Here several considerations need to be taken into account. First, the feature set needs to be optimised to minimise the noise produced from uninformative features. Second, the parameters of the SVM itself need to be tuned. As TextHunter aims to provide a general IE methodology for simple concepts, one of its features is to tune these aspects automatically.

The TextHunter pipeline always uses token lemmas (produced by the GATE Morphological Analyser), and token POS tags (produced by the GATE ANNIE POS Tagger) as features, supplied as a 'bag-of-words'. For the identified NE produced in Step 1, the NE tokens undergo syntactic processing for negation, subject and temporality detection via the ConText algorithm [130] to create additional features. No feature selection is implemented.

A configurable x-fold cross validation (default 10) is executed to produce precision and recall statistics. This result is then stored, and the SVM parameters are changed, by

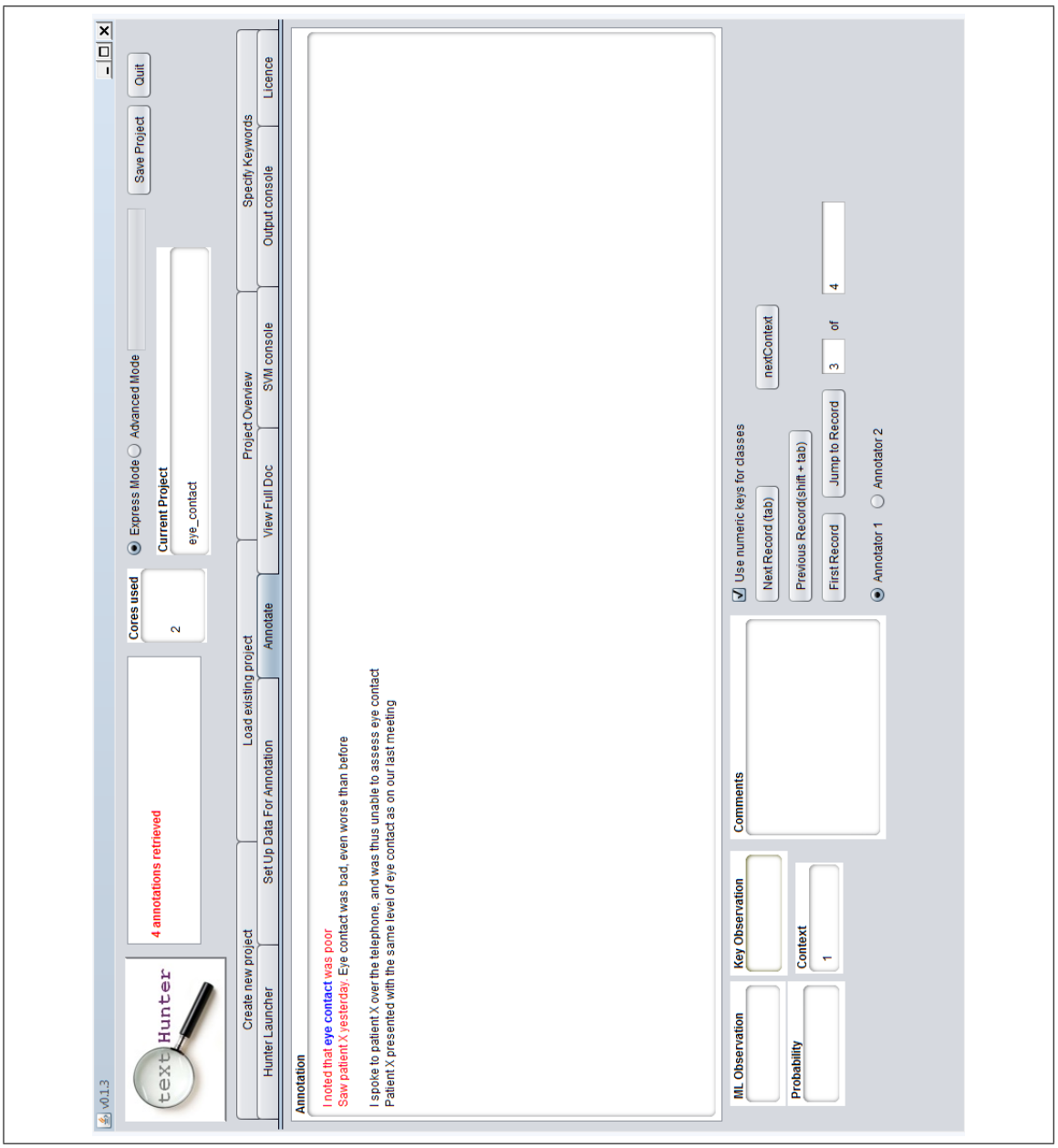


Figure 4.2: TextHunter annotation interface.

varying the cost, kernel type, degree of the kernel (if relevant), the tau parameter, and whether to include ConText features. The cross validation is then executed again. This process repeats until a predefined number of feature/parameter variants are exhausted.

The model with the best classification performance from cross validation is chosen as the 'final' model, and executed over the held out gold standard to give an estimate of final classification performance, also indicating whether the model has been overfit.

Although not implemented, it may be possible to extract additional features reflecting a richer linguistic context. For example, noun chunks and verb phrases might be obtained via dependency parsing using the Bist parser [217], which has been shown to have good classification performance on clinical text [218]. More complex features to integrate might include document metadata, such as the name of the free text source field that was used

and the job role of the author. The value of such features could be established via ablation studies. However, the inclusion of such metadata would necessitate a substantial architectural overhaul of the TextHunter application, as the Batch Learning plugin available in Gate does not easily accommodate such non-linguisitic data.

4.3.4 Step 3.ii Active Learning

Human annotation is a labour intensive process, with modern NLP systems often trained on tens or even hundreds of thousands of human labelled examples [219, 220]. Such corpora a often generated via a crowd sourcing methodolody, whereby data and annotation tools can be distributed to workers all over the world via systems such as Amazon's Mechanical Turk. Such corpora generally reflect articles intended for general consumption (such as news stories or Wikipedia articles), and the annotation process is based upon the assumption that the annotator can interpret the data presented to them without a specialised education. However, it has also been shown that crowd sourcing annotations for specialised domains such as the biomedical literature can lead to cheap, high quality annotations [221]. Nevertheless, such strategies cannot assist the annotation process where access to data is highly privileged, as is generally the case with clinical data.

Active learning is an approach to the human annotation of data that seeks to maximise the ratio of volume of data annotated to predictive performance of a model [222,223]. The general principal involves some initial training of an algorithm based upon a small 'seed' set of labelled data. The model is then applied to a much larger set of unlabelled data, whereupon it prioritises the labelling of additional instances according to some criteria. The goal of this methodology is to reduce the likelihood that a human annotator will waste time labelling uninformative instances that do not change the predictions a model makes.

To illustrate via a toy example, a dataset might contain sentences of only six tokens or less, that indicate whether a patient has poverty of speech or not. A classifier makes it's decision based on a simple vocabulary of seven token: ['the', 'patient','has','no','poverty',''of','speech']. If the classifier detects the token 'no' and any of the other tokens, it labels the instance as 'negative'. If the classifier detects the other tokens apart from the token 'no' in it's vocabulary, it labels the instance as 'positive'. To implement a trivial example of active learning, the classifier might implement some logic to identify instances containing information that isn't present in its model. For example, if classifier detects a token not in its vocabulary, it might prioritise this instance for annotation, as the unseen token may be critical to determining which label should be assigned. Without this active learning step, a user conducting annotation work over randomly selected examples might observe an instance that adds no information to the classifier, such as the sequence of tokens: ['the',

‘patient’,‘has’,‘poverty’,‘’of’,‘speech’].

In practice, approaches to active learning are much more sophisticated, often leveraging latent statistical information derived from the underlying model in order to determine how to rank annotations for further annotation work. In SVMs, Schohn and Cohn [224] devised a simple heuristic based upon selecting examples that are in close proximity to the hyperplane (see section 2.3.4). The intuition of this approach is that the SVM algorithm attempts to find the hyperplane that best separates the labelled instances as they are represented in vector space - i.e. the distance between the nearest points of two classes are maximised. By prioritising these instances for annotation, the hyperplane is guaranteed to be modified in the next iteration of it’s calculation. Via empirical experimentation, Schohn and Cohn discovered that this method of active learning via ‘maximal margin’ not only reduced the amount of training data required to produce an optimal solution, but also performed better than a model trained with *all* instances labelled.

TextHunter employs this form of active learning. If the model classification performance doesn’t meet expectations from an initial round of ‘seed’ annotations, the user can attempt to improve it by adding more training data. This can be done either by simply annotating more randomly selected instances, or by using an active learning inspired approach as follows. For the latter, first, the intermediate, under-performing model is applied over all instances discovered in stage 1. During this step, the SVM confidence (distance to the hyperplane) is captured and persisted to the database table. These are presented to the user in ascending order of hyperplane distance, thus prioritising instances for annotation. In an ideal theoretical scenario, the user would label the instance closest to the hyperplane and repeat. However, as noted by Lewis et al [225]), training an SVM and reclassifying instances can be computationally expensive and not practical. Therefore, a user typically labels a batch of instances close to the hyperplane before the cycle is repeated. Once the sample of newly labelled instances are added into the training data, step 3.i is repeated and the new result can be assessed. If results fail to improve after a user defined number of active learning iterations, the user might reasonably assume that their use case is too complex to be handled by the TextHunter pipeline.

More sophisticated approaches to active learning have also been devised. Culotta and McCallum developed an annotation framework for prioritising annotation that calculated not only the information content of a given instance, but the difficulty in annotating it [226]. Here, the authors considered a multi-label annotation task, wherein the annotation process is optimised by offering a selection of the most likely labels predicted by the classifier, instead of the user having to select the correct label from a potentially long list. In addition, their framework leverage’s correct predictions made by a model, negating the

need for users to completely re-annotate instances that a model has partially predicted correctly.

Working on the impracticalities of retraining and re-applying a classifier after every new example, Brinker [227] proposed a more sophisticated approach to labelling batches of prioritised instances between iterations than simply selecting the top n instances closest to the hyperplane. Here, Brinker suggested that batches could be optimised by incorporating a measure of information diversity into the ranking of instances. Brinker identified that beyond the first instance selected via the standard distance metric described above, there was no guarantee that the hyperplane would be affected - intuitively to the task of text classification, the second sentence in a prioritised list of instances might be identical to the first, or not different enough to have much effect on overall classification performance. The proposed solution integrated information about the distance between unlabelled instances close to the hyperplane in the version space, when suggesting a priority batch for labelling.

Such approaches have been employed in clinical/biomedical text classification. Figueroa et al [228] incorporated informativeness (distance to hyperplane), diversity and a third metric, representativeness into their active learning methodology for NER tasks. Here, representativeness is intended to prevent outliers in the underlying data from being selected for annotation, by ensuring that it is similar to other training examples. To calculate this, they use cosine distance of a given word based upon it's feature vector of as projected in the SVM version space. In evaluation upon the GENIA corpus, they found that employing such strategies lead to a reduction of up to 40% of the required training data compared to the random selection of instances alone.

Tsuruoka et al [229] proposed a strategy to overcome the sampling bias associated with active learning in NER tasks. Here, the authors recognise the implicit bias that active learning introduces into a training data set. Since the bias is dependent on the algorithm used in selecting instances, data collected in such a way may be problematic if it is to be re-purposed for a use case other than the optimisation of the classification performance of the original algorithm (for example, for enriching an existing corpora labelled in an unbiased way with additional information). The authors propose a methodology by which a CRF classifier is trained on a portion of sentence in a corpus to find named entities of interest. This is then applied over the whole corpus, and the sentences deemed likely to contain entities are selected for annotation. This process is then repeated until a desired level of coverage of the corpus is reached. Although the method implies that some named entities will be missed due to misclassifications of the CRF, their empirical evidence suggests that the annotation cost of the GENIA corpus could be halved at the price of accepting a missing annotation rate of only 1.0%

Finally, in the clinical space, Figueroa et al [230] experimented with the effect of diversity and SVM hyperplane distance based active learning approaches. Here, the authors attempted to quantify the value of active learning on text classification problem not unlike those evaluated in the TextHunter paper described above, looking at depression and smoking status/history. They found distance and distance combined with diversity to be valuable in reducing the volume of data required for a given level of classification performance - a result consistent with my results obtained via TextHunter. Notably, they did not observe any value in incorporating a diversity metric in isolation, in instance selection - a result they attribute to the low diversity of clinical concepts.

4.3.5 Step 3.iii - Confidence Evaluation

Upon completion of the model, the final step is to select an appropriate confidence margin as described in the TextHunter paper. The model building process produces a tab delimited file of the confidence probabilities for all instances in the gold standard. Listing 4.2 describes an R script than can be used against this file, to produce an ROC analysis, estimating the recall at 95%, 90% and 85% precision.

Listing 4.2: ROC analysis R script

```
mydata = read.delim("<path_to_probabilities.tsv>", header = TRUE, row.names = 1
    ,sep = "\t")
mydata <- mydata[,colSums(is.na(mydata))<nrow(mydata)]

mydata2 <-mydata[c(2:2)]


projectNames<-colnames(mydata2)
projectColors<-list()
projectColors[[1]] = 1
projectColors[[2]] = 2
projectColors[[3]] = 3
projectColors[[4]] = 4
projectColors[[5]] = 5
projectColors[[6]] = 6
projectColors[[7]] = 7
projectColors[[8]] = 8
#projectColors[[9]] = 9
```

```
library(ROCR)


#get model names (trim df)
#titlenames <- colnames(mydata[c(-1,-(length(mydata)))])
titlenames<-"X14"
titlenames<-projectNames

mydata$observationclean[mydata$observation=="positive"] <- TRUE
mydata$observationclean[mydata$observation=="negative"] <- FALSE
mydata$observationclean[mydata$observation=="unknown"] <- FALSE
mydata$observationclean[mydata$observation=="form"] <- FALSE


pred.obs<-list()
roc.obs<-list()
prec.rec.obs<-list()
cols<-list()

for (i in seq(titlenames)) {
   pred.obs[[i]] = prediction( mydata[,i+1], mydata$observationclean)
   roc.obs[[i]] = performance(pred.obs[[i]], "tpr","fpr")
   prec.rec.obs[[i]] = performance(pred.obs[[i]], "prec","rec")
}


rocaTest<- function(mydata,projectColors, x,bestModel,addToPlot){
library(ROCR)


#get model names (trim df)
#titlenames <- colnames(mydata[c(-1,-(length(mydata)))])
titlenames <- bestModel

mydata$observationclean[mydata$observation=="positive"] <- TRUE
mydata$observationclean[mydata$observation=="negative"] <- FALSE
mydata$observationclean[mydata$observation=="unknown"] <- FALSE
mydata$observationclean[mydata$observation=="form"] <- FALSE
```

```
pred.obs<-list()
roc.obs<-list()
prec.rec.obs<-list()
cols<-list()

for (i in seq(titlenames)) {
   pred.obs[[i]] = prediction( mydata[,i+1], mydata$observationclean)
   roc.obs[[i]] = performance(pred.obs[[i]], "tpr","fpr")
   prec.rec.obs[[i]] = performance(pred.obs[[i]], "prec","rec")
}



precCutoff <- 0.90
precMax <- 0.0
recMax <- 0.0
cutoff <-0.0
besti = 0

for (i in (seq(titlenames))-1) {
   for (ii in (seq(prec.rec.obs[[i+1]]@y.values[[1]])-1)){
      if(isTRUE(prec.rec.obs[[i+1]]@y.values[[1]][ii] > precCutoff &&
          isTRUE(prec.rec.obs[[i+1]]@x.values[[1]][ii] > recMax ))){
         precMax <-prec.rec.obs[[i+1]]@y.values[[1]][ii]
          recMax <-prec.rec.obs[[i+1]]@x.values[[1]][ii]
          cutoff <-prec.rec.obs[[i+1]]@alpha.values[[1]][ii]
          modelName <- titlenames[i+1]
          besti = i+1
      }
   }
}
print(paste("90% Project Name ", projectNames[x],"Model number = " ,modelName,
    "Prec = ", precMax, "Rec = ", recMax, "cutoff = ", cutoff))

precCutoff <- 0.95
precMax <- 0.0
recMax <- 0.0
cutoff <-0.0
besti = 0

for (i in (seq(titlenames))-1) {
```

```
   for (ii in (seq(prec.rec.obs[[i+1]]@y.values[[1]])-1)){
      if(isTRUE(prec.rec.obs[[i+1]]@y.values[[1]][ii] > precCutoff &&
          isTRUE(prec.rec.obs[[i+1]]@x.values[[1]][ii] > recMax ))){
         precMax <-prec.rec.obs[[i+1]]@y.values[[1]][ii]
          recMax <-prec.rec.obs[[i+1]]@x.values[[1]][ii]
          cutoff <-prec.rec.obs[[i+1]]@alpha.values[[1]][ii]
          modelName <- titlenames[i+1]
          besti = i+1
      }
   }
}
print(paste("95% Project Name ", modelName,"Model number = " ,modelName, "Prec =
    ", precMax, "Rec = ", recMax,"Cutoff = ",cutoff ))

precCutoff <- 0.85
precMax <- 0.0
recMax <- 0.0
cutoff <-0.0
besti = 0

for (i in (seq(titlenames))-1) {
   for (ii in (seq(prec.rec.obs[[i+1]]@y.values[[1]])-1)){
      if(isTRUE(prec.rec.obs[[i+1]]@y.values[[1]][ii] > precCutoff &&
          isTRUE(prec.rec.obs[[i+1]]@x.values[[1]][ii] > recMax ))){
         precMax <-prec.rec.obs[[i+1]]@y.values[[1]][ii]
          recMax <-prec.rec.obs[[i+1]]@x.values[[1]][ii]
          cutoff <-prec.rec.obs[[i+1]]@alpha.values[[1]][ii]
          modelName <- titlenames[i+1]
          besti = i+1
      }
   }
}
print(paste("85% Project Name ", modelName,"Model number = " ,modelName, "Prec =
    ", precMax, "Rec = ", recMax,"Cutoff = ",cutoff ))

   if(addToPlot){
      plot(roc.obs[[besti]], col= 1, add=TRUE)
   }else{
      par(mfrow=c(1,1))
      plot(roc.obs[[besti]], col= 1)
```

```
  }
}

addToPlot <-FALSE
rocaTest(mydata,projectColors,2,titlenames,addToPlot)

 #legend("bottomright", legend=projectNames, pch = 15, col=unlist(projectColors)
     )
```

4.3.6 Step 4 - Model Application

Once a model with satisfactory classification performance (as determined by the user's use case) is obtained, it can be applied over all unannotated instances, with the results persisted to the target database. From here, they can be combined with other EHR data for dataset construction via standard SQL query syntax. Scaling is achieved vertically, using a multi-threaded approach within a single JVM.

4.3.7 Limitations/Further Work

Our initial investigations on negative symptoms suggested that the best classification performance was achieved via combining concept specific rules and concept specific training data to produce a hybrid classification approach. However, as previously discussed, producing training data can be made substantially easier than producing rules. With TextHunter, we sought to leverage this feature, while still retaining some rule based syntactic processing in the form of the ConText algorithm. Our results suggest that this approach provides good classification performance on simple constructs, but works less well on more complex ones. For instance, a relatively poor f1 score was observed in the ethnicity case study. Here, the number of contexts in which a putative mention of ethnicity can occur are far greater than the contexts in which mentions of 'hallucinations' can occur, such as the use of demonyms (e.g. "a 52 year old Ghanan man visited today, with his half-Ghanan, half English son", or "a 21 year old Swiss woman is pursuing English Literature studies at University College London, and is in the process of obtaining her British citizenship"). Clearly, such constructs would require additional feature representation than what is available from our BOW methodology in order for a classifier to have access to the necessary information. Such features might be provided by the use of dependency parsers as described above.

As discussed in section 2.3.4, several studies have suggested that CRF sequence style classifiers may outperform SVMs in IE tasks. I investigated using the MALLET library

[231] as an alternative methodology to SVMs in TextHunter. However, at the time that TextHunter was developed, this library suffered thread safety issues, meaning that it would not be suitable for addressing our practical requirements of scalability. This issue appears to have since been addressed, and may be a consideration for further refinement of the TextHunter application.

4.4 The CRIS-CODE project

Following the completion of the main features of the main features of TextHunter, it was possible to commence work on collecting comprehensive symptomatology and subsequently profiling the patient population at SLAM. This work is encapsulated by the resulting publication (author contributions are listed in the paper):

BMJ Open

Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project

Richard G Jackson,[1] Rashmi Patel,[1] Nishamali Jayatilleke,[1] Anna Kolliakou,[1] Michael Ball,[1] Genevieve Gorrell,[2] Angus Roberts,[2] Richard J Dobson,[1] Robert Stewart[1]

To cite: Jackson RG, Patel R, Jayatilleke N, *et al*. Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ Open* 2017;**6**: e012012. doi:10.1136/bmjopen-2016-012012

► Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2016-012012).

Received 23 March 2016
Revised 11 August 2016
Accepted 4 October 2016

[1]Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK
[2]Department of Computer Science, University of Sheffield, Sheffield, UK

Correspondence to
Richard G Jackson;
richgjackson@gmail.com

ABSTRACT

Objectives: We sought to use natural language processing to develop a suite of language models to capture key symptoms of severe mental illness (SMI) from clinical text, to facilitate the secondary use of mental healthcare data in research.

Design: Development and validation of information extraction applications for ascertaining symptoms of SMI in routine mental health records using the Clinical Record Interactive Search (CRIS) data resource; description of their distribution in a corpus of discharge summaries.

Setting: Electronic records from a large mental healthcare provider serving a geographic catchment of 1.2 million residents in four boroughs of south London, UK.

Participants: The distribution of derived symptoms was described in 23 128 discharge summaries from 7962 patients who had received an SMI diagnosis, and 13 496 discharge summaries from 7575 patients who had received a non-SMI diagnosis.

Outcome measures: Fifty SMI symptoms were identified by a team of psychiatrists for extraction based on salience and linguistic consistency in records, broadly categorised under positive, negative, disorganisation, manic and catatonic subgroups. Text models for each symptom were generated using the TextHunter tool and the CRIS database.

Results: We extracted data for 46 symptoms with a median F1 score of 0.88. Four symptom models performed poorly and were excluded. From the corpus of discharge summaries, it was possible to extract symptomatology in 87% of patients with SMI and 60% of patients with non-SMI diagnosis.

Conclusions: This work demonstrates the possibility of automatically extracting a broad range of SMI symptoms from English text discharge summaries for patients with an SMI diagnosis. Descriptive data also indicated that most symptoms cut across diagnoses, rather than being restricted to particular groups.

Strengths and limitations of this study

- The number and diversity of symptomatology concepts that we successfully modelled indicates that this task is suitable for natural language processing.
- The large number of records in the Clinical Record Interactive Search database gives insight into the reporting realities of symptomatology in a typical UK National Health Service Mental Health Trust for individuals who have received an International Classification of Diseases, Tenth Revision, severe mental illness (SMI) diagnosis.
- Our negative control group suggests a wide under-reporting of SMI symptoms in patients who have not received an SMI diagnosis, although our models were not validated in this group and such patients may have later received an SMI diagnosis after our analysis was concluded.
- Similarly, our models were validated on English text from a single UK site—the models may not generalise across different institutions and geographic/medical dialects.
- We did not attempt to resolve temporal aspects of symptomatology in this study, which will be necessary for future predictive modelling approaches.

INTRODUCTION

EHRs in health research

Electronic health records (EHRs) are recognised as a valuable source of data to support a wide range of secondary informatics use cases, such as decision support, observational research and business intelligence.[1] With appropriate handling, EHRs may be able to overcome the cost barriers to generating sufficient data for addressing complex questions that would be out of reach for more

conventional patient recruitment protocols.[2–4] However, the use of EHRs in this way is known to create a range of new issues that need to be addressed before the data can be considered of sufficient quality suitable for research.[5]

Symptomatology of severe mental illness

In mental health research and clinical practice, it is often argued that the symptoms expressed by a patient in the course of their illness represent a more useful description of the disorder and indications for intervention than the concept of a diagnosis.[6 7] While common conditions in mental health are represented in classification taxonomies such as the International Classification of Diseases (ICD) and Diagnostic and Statistical Manual (DSM) systems, generally speaking, it is the symptomatology of a condition that is used by clinicians to determine an appropriate treatment plan. This is due to the broad symptomatic manifestations of mental disorders, in the sense that, at a given time, a patient assigned a diagnosis (such as schizophrenia) can present with all, many or very few of the symptoms associated with the condition. This is particularly pertinent to clinical practice where diagnoses are not necessarily assigned using research criteria. The problems of diagnostic semantics are especially apparent in severe mental illness (SMI; schizophrenia, schizoaffective disorder and bipolar disorder). Here, the controversy is compounded by the high frequency of mental health comorbidities and shortcomings in our current understanding of the biological underpinnings of mental disorders, which in turn limit our ability to subclassify the conditions. For example, Van Os *et al*[8] suggest that there are overlapping genetic, neurobiological and clinical features between different categories of mental disorder, and Insel *et al*[9] suggest that within each diagnostic category there is a considerable degree of heterogeneity and that the diagnostic category in itself provides little information about future clinical outcomes. In addition, the lack of genetic and other objective tests for many mental disorders has led to a requirement for detailed, interpersonal observation of patients, cumulating in pragmatic symptomatology-based assessments.[10–14] Information on specific symptoms is typically recorded in unstructured parts of the EHR,[15] and the incorporation of structured instruments for recording symptoms has not so far proved feasible in routine clinical practice outside specialist services. Hence, the free text portion of the mental health EHR contains a potentially vast and complex tapestry of clinical information which to date has been effectively 'invisible' when it comes to the generation of data for administration, business intelligence, research or clinical evaluation.

Such a situation thus represents a quandary for mental health informaticians and clinical researchers alike. A common task in health research is to group patients with similar conditions into appropriate cohorts, which will almost inevitably require ascertaining common factors pertinent to their disorder.[16–18] Diagnoses form semantically convenient units, although the usefulness may be disputed and/or lacking in granularity. Symptomatology may offer more objective, relevant groupings but the data may be locked in unstructured free text, presenting unique data extraction problems.

Natural language processing and information extraction

Natural language processing (NLP) and its subdiscipline of Information Extraction (IE) are commonly employed within clinical records to process large quantities of unstructured (human authored) text and return structured information about its meaning.[19–21] Medical entities frequently targeted include medications, diagnoses, smoking status and other factors influencing risk, course or outcome for disorders of interest.[21 22] A large number of tools and frameworks exist for general purpose information extraction from clinical dictionaries, such as cTAKES,[22] NOBLE[23] and MedLee.[24] However, there has been little application of NLP techniques in mental healthcare data despite the volumes of text-based information contained here, and even less on ascertaining symptomatology. Here, we introduce the CRIS-CODE project, which has the long-term objective of offering comprehensive NLP models for mental health constructs. The focus of the initial programme of work described here was to develop sentence classification models for a substantial range of SMI symptomatology, to allow automatic extraction for many of the most informative symptoms from the patient narrative. It is envisaged that the outcomes will support a range of future research and clinical applications.

MATERIALS AND METHODS

Corpus selection and preprocessing: the South London and Maudsley Mental Health Case Register

The South London and Maudsley NHS Foundation Trust (SLaM) is one of the largest mental healthcare organisations in Europe, and provides mental health services to 1.2 million residents in its geographic catchment of four south London boroughs (Lambeth, Southwark, Lewisham and Croydon), in addition to national specialist services. SLaM adopted fully EHRs for all its services during 2006, importing legacy data from older systems during the process of assembly. In 2007–08, the Clinical Record Interactive Search (CRIS) application was developed with funding from the British National Institute for Health Research.[25] CRIS generates a research database consisting of a pseudonymised version of SLaM's EHR system: currently containing de-identified patient records on more than 250 000 patients and over 3.5 million documents in common word processor formats. Since its development, the data contained have been substantially enhanced through external linkages and NLP.[26] Patient consent was not required for this retrospective study.

Definitions of SMI symptoms

A keyword lexicon of SMI symptoms was defined by a team of psychiatrists, based on pragmatic criteria. First, the potential salience of symptoms for research applications was considered, particularly their incorporation in symptom scales in common clinical use, such as the Positive and Negative Symptoms Scale (PANSS)[13] and Young Mania Rating Scale (YMRS)[27] which were used as templates for guidance. Second, the language used in routine clinical records was taken into consideration in choosing symptoms, focusing particularly on those which were likely to be recorded in the most consistent and tractable language, based on clinical experience. Third, we sought a priori to extract sufficient numbers of symptom types to generate scales for further evaluation within the following five domains: (1) positive symptoms; (2) negative symptoms; (3) disorganisation symptoms; (4) manic symptoms and (5) catatonic symptoms. The first four of these followed the findings of Demjaha *et al*,[28] although we had not at this point attempted to define depressive symptoms. Catatonic symptoms were further added to improve consistency with the study of Cuesta and Peralta,[29] and as a symptom group of interest, which is often not adequately captured in dimensional studies because of its relative rarity in recruited clinical samples.

We defined the NLP task as a sentence classification problem, with a classifiable instance as a sentence containing a symptom keyword or the general constructs of 'negative symptoms' or 'catatonic syndrome' (referring to groups 2 and 5 above). In addition to the keywords, clinically relevant modifier terms were also defined for some concepts, in order to produce subclassifications of symptoms where appropriate (table 1). If a modifier term was detected within eight words of a keyword, the modifier was deemed to be a possible relation. We further specified that modifiers could be 'mandatory' (meaning a modifier was required to be present for our definition of an instance to be met), or 'optional' (meaning only the keyword needed to be present for our instance definition to be met) (table 2). Regarding potential biases that might result from missing synonyms outside of our selected keywords, we did not consider this to be a significant problem. Clinical staff receive substantial training about how to document symptomatology in specific ways, in order to differentiate between a clinical opinion ('the patient exhibited apathy') and a non-clinical opinion ('the patient expressed indifference towards their treatment today'), and therefore chose our keywords in line with the standard methods of symptom documentation to avoid uncertainty in the authors intent. Similarly, our objective was to identify clinician-assigned constructs, rather than attempt to classify descriptions of experiences—for example, identifying the recorded assignment of 'hallucination' as a symptom, rather than the description of the person's perceptual disturbance; identifying the recording of 'delusion' rather than the description of the false belief.

Information extraction with TextHunter

TextHunter is an NLP information extraction suite developed jointly by SLaM and the Institute of Psychiatry, Psychology & Neuroscience at King's College London.[30] Its principle purpose is to provide an interface to accomplish three tasks required to extract concepts from free text:

1. find instances of a concept in a database of documents using regular expression style matching of keywords;
2. provide an efficient interface to allow human annotators to label a portion of the sentences containing the concept instances in order to develop a gold standard and training corpora;
3. attempt to construct an appropriate support vector machine (SVM) language model of the concept, and validate it with the gold standard corpus.

Briefly, TextHunter is built around the ConText algorithm[31] and the GATE framework Batch Learning plugin, a machine learning framework which in turn uses the LibSVM java library.[32] A SVM is a machine learning methodology that maps the features of human labelled input training data instances into vector space. Within this space, a learning algorithm is applied to construct a hyperplane, which attempts to accurately differentiate the different training instances based on their labels. Once this hyperplane is 'learnt', the model can be applied to new, unseen instances to predict the label that should be assigned. TextHunter uses bag-of-words features such as keywords, surrounding word tokens and part-of-speech tags in conjunction with knowledge engineering features generated from ConText to build a sentence classifier. A full description of its workings is described in ref. 30. In this analysis, we used V.3.0.6 of TextHunter.

Annotation of SMI symptom concepts

In order to produce annotation guidelines to ensure consistent, high-quality gold standard and training data, we developed annotation guidelines based around internal, iterative discussions. Generally, we defined a relevant instance as a mention of a symptom observed in a patient, without a grammatical negation. Owing to the large numbers of concepts addressed by this work, it was only feasible to double annotate 15 of the concepts to derive interannotator agreement (IAA) statistics. This was completed by either two psychiatrists or a psychiatrist and a trained research worker familiar with the construct.

To optimise the performance of the language models for the SMI cohort, we enriched our training corpus by selecting any text occurrence in CRIS (irrespective of the document type), relating to a patient who had received an SMI diagnosis, defined as schizophrenia (ICD-10 code F20x), schizoaffective disorder (F25x) or bipolar disorder (F31x). This diagnosis information came from structured fields in the source EHR, which are completed by clinicians during the normal course of

Table 1 Symptom instance definitions

SMI concept	Keyword strings	Modifier strings	Lax or strict modifiers	SNOMED-CT (SCTID)†
Aggression	aggress*			61372001
Agitation	agitat*			106126000
Anhedonia	anhedon*			28669007
Apathy	apath*			20602000
Arousal	arous*			(none)
Blunted or flat affect	Affect	blunt*, flat*, restrict*	Optional	6140007/932006/39370001
Catalepsy	catalep*			247917007
Catatonic syndrome	catatoni*			247917007
Circumstantial speech	circumstan*			18343006
Deficient abstract thinking	Concrete			71573006
Delusions	delusion*			2073000
Derailment of speech	derail*			65135009
Diminished eye contact	eye contact			412786000
Disturbed sleep	Sleep	not, poor*, interrupt*, nightmare*, disturb*, inadequat*, disorder*, prevent*, stop*, problem*, difficult*, reduced*, less*, impair*, erratic*, unable*, worse*, depriv*	Optional	26677001
Echolalia	Echolalia			64712007
Echopraxia	Echopraxia			33184005
Elation	elat*			34822003
Elevated mood	Mood	elevat*	Mandatory	81548002
Emotional withdrawal	withdraw*			247755007
Euphoria	euphor*			85949006
Flight of ideas	flight of idea			28810003
Formal thought disorder	Ftd thought disorder			41591006
Grandiosity	grandios*			247783009
Hallucinations	hallucinate*	audit*, visual*, olfact*, tactil*, third person, first person, 3rd person, 1st person,	Optional	45150006/64269007/39672001/66609003/277533007/
Hostility	hostil*			79351003
Immobility	immobil*			404975000
Insomnia	insom*			193462001
Irritability	irritabl*			55929007
Loosening of associations	associat*			55346003
Loss of coherence	coheren*			284596004
Low mood	Mood			366979004
Mannerisms	Mannerism*			248026005
Mutism	Mute Mutism			88052002
Negative syndrome	negative symptom*			(none)
Paranoia	paranoi*			191667009
Persecutory ideation	persecu*			216004
Perseverance	persever*			44515000
Poor motivation	motivat*			26413003
Poor rapport	Rapport			710497003
Posturing	postur*			271694000
Poverty of speech	speech*	poverty*, impoverish*	Mandatory	72123004
Poverty of thought	poverty of thought			56435009
Pressured speech	speech*	pressure*	Mandatory	53890003
Rigidity	rigid*			311535006
Social withdrawal	withdraw*	social*	Mandatory	105411000
Stereotypy	stereotyp*			84328007

Continued

Table 1 Continued

SMI concept	Keyword strings	Modifier strings	Lax or strict modifiers	SNOMED-CT (SCTID)†
Stupor	stupor*			89458003
Tangential speech	tangent*			74396008
Thought block	though* block			2899008
Waxy flexibility	Waxy			13052006

†Best matches in SNOMED-CT, UK-edition v20160401.
SMI, severe mental illness; SNOMED, Systematized Nomenclature of Medicine; SNOMED-CT (SCTID), Systematized Nomenclature of Medicine—Clinical Terms Identifier.

care by means of selecting an appropriate ICD-10 code; these were supplemented by a separate NLP application[33 34] which returns searchable text strings associated with diagnostic statements in text fields. In UK NHS Mental Health Trusts, recording of diagnosis is effectively mandatory, but recorded diagnoses themselves have no financial implications for trusts (eg, are not used for any billing purposes).

An independent set of gold standard data were also created for each symptom to assess the performance of each model. This was derived in the same manner as the training data.

For training and gold standard data, a relevant instance of a symptom was labelled as 'positive', (such as 'the patient had poverty of speech') whereas irrelevant or negated instances (such as 'today I examined the patient for poverty of speech…' or 'the patient did not have poverty of speech') were labelled as 'negative' to create a binary classification problem (for the special case of the 'negative symptoms' construct, this was annotated as positive when described as present (eg, 'he experiences severe negative symptoms') and negative when absent (eg, 'there was no evidence of negative symptoms')). The training data were then used in 10-fold cross validation to estimate the optimal SVM model parameters using the features provided by TextHunter (see above). An instance was considered correctly classified if the sentence containing the human label of 'positive' or 'negative' and symptom type matched the model-generated label and symptom type. Subclassifications of classes based on any modifiers that were present were not evaluated in this work. Finally, we validated the optimised models against our gold standard data. We arbitrarily decided that the gold standard for each concept should contain a minimum of 100 'positive' mentions, in order to derive precision, recall and F1 measures for the 'positive' class.

Owing to the tendency of a given set of clinical notes to repeat certain pieces of information over time, EHRs offer multiple opportunities to bolster recall (eg, via the reassessment of symptoms across multiple visits). For this reason, we favoured precision over recall as the more desirable performance metric. We applied SVM confidence margin filters to increase precision where acceptable losses to recall were possible. If performance was deemed to be poor, we attempted to improve the model by adding further training data, in some cases using TextHunter's active learning capability. In addition, we evaluated the accurate identification of the negation status of each symptom between TextHunter + ConText rules versus the ConText negation feature in isolation.

Descriptive statistics of SMI distribution among SMI and non-SMI cohorts

A cohort of 18 761 patients was selected from CRIS, dating from the inception of electronic records in SLaM in 2006 to November 2014, all of whom had received an SMI diagnosis as defined above at any point during that period. For a negative control, we also selected a cohort of 57 999 patients that had received a non-SMI diagnosis, defined as the assignment of an ICD-10 code of F32 (depressive episode), F33 (recurrent depressive disorder, F40–F48 (neurotic, stress-related and somatoform disorders) or F60 (personality disorder) in the same period. F32.3 (severe depressive episode with psychotic symptom) and F33.3 (recurrent depressive disorder, current episode severe with psychotic symptoms) were excluded from the non-SMI cohort so as to not overlap

Table 2 Examples of instances

Type	Keyword	Modifier	Example
Mandatory	Speech	pov*	There was some *poverty* of *speech* and content of thought.
Optional	hallucinat*	Audit*	For past 1 week has been having *auditory* command *hallucinations* telling him to kill himself and also suicidal ideation.
Optional	hallucinat*		These *hallucinations* are sometimes in a kind of shadow form shaped like a man I call 'David' and 'James'.
None	Rapport		When she was last seen at her CPA on XXXXX by Specialist Registrar Dr XXXXX, ZZZZZ presented as well kempt with good eye contact and *rapport*.

with our SMI group. The NLP models were applied to a corpus of documents labelled as discharge summaries linked to these cohorts, and descriptive statistics were collected from the results.

RESULTS

Interannotator agreement and model validation

In total, 50 different symptoms were chosen, and a total of 37 211 instances of symptoms were annotated from 32 767 documents to create gold standards and training data specific to each symptom. An additional 2950 instances across 15 symptoms were double annotated (table 3), yielding an average Cohen's κ of 0.83.

Across all 50 symptoms, the average count of instances per gold standard was 202. Of the 50 symptoms for which we attempted to build models, four performed poorly (loosening of associations, stereotypy, low mood and poor motivation). Two symptoms were so rare (catalepsy, echopraxia) that it was practical to annotate all detected mentions of the keywords by hand. One symptom (mutism) achieved an acceptable performance based on the mention of the symptom keyword alone. Of the remaining 43 symptoms, the hybrid model produced a precision of at least 85% in 38 symptoms, compared with 23 symptoms using the ConText negation model alone. The precision, recall and F1 metrics of each modelled symptom for individuals with an SMI diagnosis are listed in online supplementary table 1. Summary statistics aggregated across all symptoms for each approach are presented in table 4.

Analysis of discharge summaries

Of the 18 761 patients in our SMI cohort, we were able to identify at least one labelled discharge summary for a subset of 7962 patients, to generate a corpus of 23 128 discharge summaries. For the 57 999 patients in our non-SMI cohort, we identified 13 496 discharge summaries for a subset of 7575 patients. The 43 NLP models were applied to the SMI and non-SMI corpora, which returned a total of 171 523 symptoms in 17 902 (77%) summaries across 6 920 (87%) patients in the SMI cohort and 31 769 symptoms in 7 259 (54%) summaries across 4540 (60%) patients in the non-SMI cohort (when combined with additional data from the three symptoms where NLP was not necessary). For succinctness, we grouped the symptoms into five semantic types, as described in table 5. The most common types were the positive symptoms (9662 patients) and the least common were the catatonic symptoms (1363 patients) (table 6). In figures 1 and 2, we plot bar charts of the counts of unique patients exhibiting each symptom, coloured by the original ICD-10 diagnosis and symptom domains respectively. In the SMI cohort, the counts of patients exhibiting the various symptoms follow an approximately Poisson distribution, with the prevalence of each symptom ranging from very common (paranoia, 59%) to very rare (catalepsy, >1%) (figure 2). In the negative control group, appreciable counts were also

Table 3 Interannotator agreement scores

Project	Instances	Observed agreement	Cohen's κ
Catatonic syndrome	232	0.88	0.65
Diminished eye contact	362	1.00	1.00
Echolalia	98	0.96	0.89
Echopraxia	93	0.99	0.98
Elation	299	0.95	0.87
Euphoria	318	0.88	0.75
Grandiosity	293	0.94	0.84
Hallucinations	137	0.91	0.81
Immobility	98	0.90	0.79
Insomnia	291	0.93	0.69
Irritability	97	0.89	0.65
Mannerisms	89	0.89	0.73
Perseverance	99	0.97	0.93
Stupor	98	0.92	0.82
Waxy flexibility	135	0.95	0.89

Table 4 Comparison of the hybrid approach and context alone across all symptoms (excluding catalepsy, echopraxia and mutism in SMI cohort)

Statistic	Model	P%	R%	F1
Mean	ConText + ML	83	78	0.80
	ConText	71	97	0.79
Median	ConText + ML	90	85	0.88
	ConText	84	98	0.91

SMI, severe mental illness.

Table 5 Symptom groupings

Domain	Symptoms
Positive	Agitation, aggression, arousal, hostility, delusions, hallucinations, paranoia, persecution
Negative	Diminished eye contact, blunted or flat affect, emotional withdrawal, social withdrawal, abstract thinking, poor rapport, apathy, anhedonia, poverty of speech, poverty of thought, negative syndrome
Disorganisation	Circumstantial speech, reduced coherence, formal thought disorder, thought block, tangential speech, derailment, flight of ideas
Manic	Elevated mood, disturbed sleep, insomnia, euphoria, pressured speech, irritability, elation, grandiosity
Catatonic	Mannerism, rigidity, posturing, perseverance, stupor, waxy flexibility, immobility, echolalia, mutism, catalepsy, echopraxia

Table 6 Counts of patients by symptom groups and ICD-10 diagnosis

Diagnosis	Catatonic	Disorganisation	Manic	Negative	Positive
F20—Schizophrenia	630	2076	2490	1903	3518
F25—Schizoaffective	71	252	370	206	432
F31—Bipolar	139	878	1316	529	1264
Multiple	268	987	1193	724	1331
Non-SMI	255	1182	3097	1984	3117
Total	1363	5375	8466	5346	9662

ICD-10, International Classification of Diseases, Tenth Revision; SMI, severe mental illness.

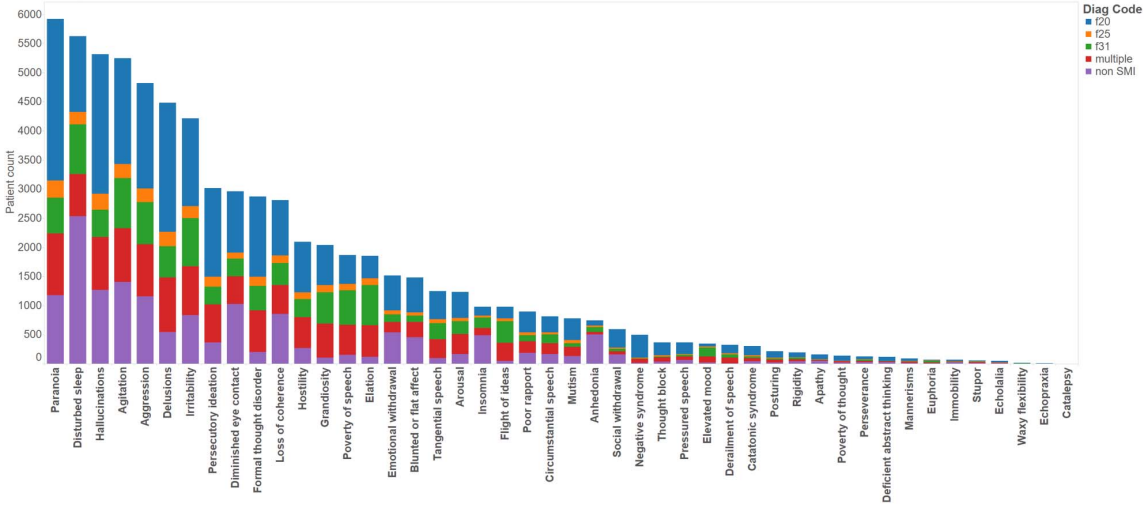


Figure 1 Distribution of symptoms by SMI ICD diagnosis. ICD, International Classification of Diseases; SMI, severe mental illness.

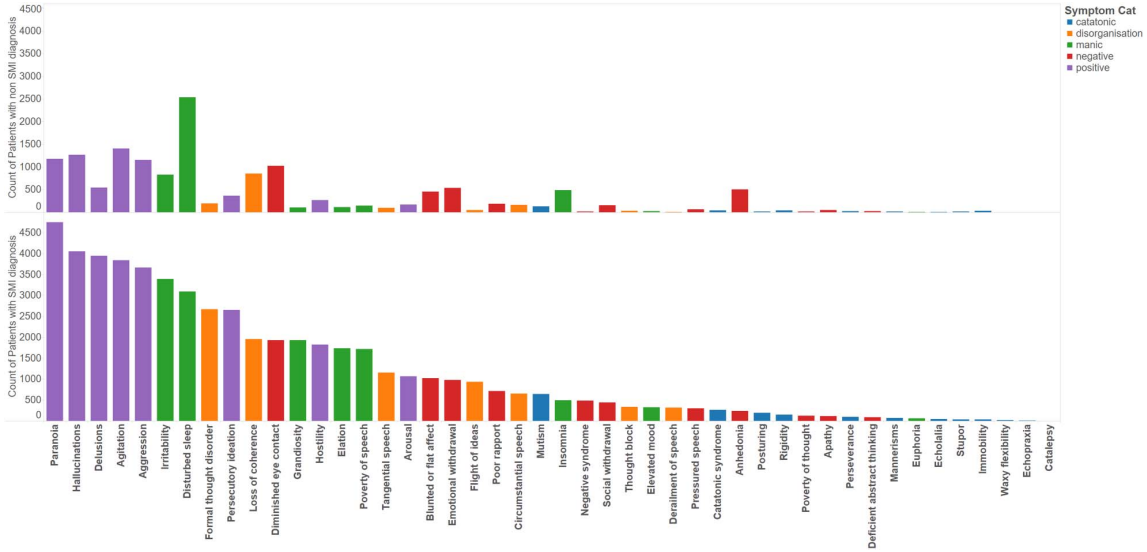


Figure 2 Distribution of symptoms by symptom classes.

observed for many of the symptoms, with disturbed sleep the most common, followed by paranoia, hallucinations, agitation, aggression, diminished eye contact and loss of coherence.

DISCUSSION

Using a large mental health EHR data resource, we were able to generate an extensive NLP-derived profiling of symptomatology in SMI, albeit limited to English language discharge summaries from patients who had received an ICD-10 SMI diagnosis. This yielded high volumes of novel information on 46 symptoms across five key domains. Comparable projects that we are aware of in mental healthcare have been the characterisation of diagnostic profiles in a Danish Psychiatric Case Register,[35] and the use of NLP-derived symptoms to assist in the diagnosis of bipolar disorder in the US EHRs.[36]

The aspiration of the 'CRIS-CODE' project is to use NLP to offer comprehensive profiling from the mental health electronic record of symptoms and of interventions, outcomes and other relevant contextual factors currently only available from text fields. Our choice of symptoms for this initial phase of information extraction was arbitrary, based on the pragmatic criteria previously stated, and not intended to be comprehensive. In addition, the categories applied to group symptoms also have to be considered as arbitrary, albeit consistent with dimensions proposed by other authors, and need further empirical evaluation. The results of the IAA exercise across 15 symptoms suggest general good agreement in our definition of symptom instances, meaning that these concepts were generally well defined and understood among clinicians. A limitation of our IAA validation approach was that we did not sample across all symptoms, as the resource overhead to train all annotators in all concepts was prohibitive. Regarding TextHunter model effectiveness, our results indicate that good information extraction performance could be achieved in the majority of SMI symptom concepts that we attempted, using the standard hybrid approach of ConText rules and machine learning offered by TextHunter. This suggests that future work to expand on this list should also be a tractable problem with this methodology. We were also able to demonstrate that the hybrid approach of combining ConText and ML generally performs favourably compared with ConText in isolation, when precision is favoured over recall, although ConText in isolation outperforms the hybrid approach if recall is favoured. As ConText was designed with medical records from the USA in mind, it is possible that the differences in medical language between British English and American English may account for the relatively low precision of ConText alone on UK medical records. However, it also conforms to the expectation that generic NLP systems for IE have limitations when applied to specific phenomena in mental health symptomatology compared with ML models trained for a specific purpose using expert clinical annotations.

In the case of some symptoms, neither the hybrid method nor ConText alone was able to deliver adequate performance. This is most likely due to the common occurrence in other contexts of the keywords used to describe instances of these symptoms, and the difficulty in disambiguating between their general use and their clinical use. For example, it is very common for a caregiver to describe a patient's 'motivation' in a variety of contexts, and differentiating a specific clinical symptom of 'poor motivation' will likely require alternative approaches. A related example might also be the variety of terms used to describe low mood, and its proximity in standard mental state examination text to statements concerning lowered or depressed *affect*—a similar but different entity ('mood' conventionally referring to a patient's reported experience of their emotional status; 'affect' to the clinician's observation of the same). It is likely that our approach of enriching the training data via selecting text from individuals with an SMI diagnosis failed to provide sufficient feature diversity for the SVMs to differentiate between relevant and irrelevant instances. Future work might address this by more detailed exploration of the common clinical language used to describe the failed concepts, in order to use knowledge engineering to derive more valuable features than a simple bag-of-words approach can yield.

An important consideration is that we were only able to identify a minimum of one symptom in 87% of patients with SMI from the corpus of documents sampled, suggesting additional recall improvements should be possible. Underestimation of symptoms may have occurred for several reasons. First, we did not specify a minimum length of treatment in our inclusion criteria, so relatively new patients with sparse documentation may not yet have any symptoms registered in their record. Second, our predilection for precision over recall in tuning our models may have reduced the probability of detection. Third, our list of symptoms was not comprehensive and may have missed some aspects of psychosis presentation—either because of different symptoms which were missed, or because of target symptoms which were described in non-standard language (eg, 'hearing voices' rather than 'auditory hallucination')—although as per our methodological reasons regarding the use of synonyms, including non-standard terms may introduce additional uncertainty as to the author's intended meaning. It is also possible that the SMI diagnosis had been first recorded at an earlier presentation and that some patients were now presenting with different sets of symptoms not currently captured (eg, people with bipolar disorder who were currently depressed, or people with previous schizophrenia currently receiving care for alcohol or drug dependence). Further in-depth exploration of text fields is warranted in the sample with no symptoms identified from the current list, to clarify the nature of symptoms and

presentations reported; such an exercise would be feasible in CRIS, but was felt to be beyond the scope of this paper. Fourth, the descriptive data were restricted to a specific corpus of documents described as discharge summaries. Discharge summaries might be considered the most 'valuable' clinical documents in NLP tasks because of their emphasis on detail and accuracy, and the tendency for institutions to encourage clinicians to use standard language in their authorship. However, it is possible that symptoms may be recorded in other areas of the record that would not have been captured by our approach. To maximise recall by including additional document types raises new questions for NLP tasks such as the importance of an author's profession and temporal aspects relating to the amount of patient/clinician contact. Finally, sufficiency of the source may be in question—for example, the CRIS database does not currently have the capacity to process scanned images of text documents (as opposed to formats such as Microsoft Word) and these images of text documents are known to make up approximately a third of all uploaded files to the clinical database. Alternatively, discharge summaries that were mislabelled as another document class at the point of upload also would not have been included in our analysis. A document classification approach may assist here.

Appreciable prevalences of many of the symptoms in the group with a non-SMI diagnosis are not unexpected, given the extent to which mental health symptoms are recognised to cross diagnostic categories—one of the factors behind CRIS-CODE's objectives. For example, sleep disturbance and diminished eye contact are common features of depressive disorder, and agitation and aggression are similarly non-specific. The common occurrence of paranoia and hallucinations would benefit from more detailed future evaluation, although might reflect early psychotic syndromes which had not yet attracted an SMI (or depressive psychosis) diagnosis, or else unrelated phenomena (eg, non-specific hallucinatory experiences accompanying sleep disturbance) or inappropriately applied terminology (eg, paranoia used to describe non-delusional hostility or suspiciousness).

CONCLUSION

The primary purpose of the developments described was to improve the depth of information available on patients with these disorders represented on healthcare datasets, as these information resources frequently contain little information beyond a diagnosis. The case for identifying symptoms of SMI as a source of data for mental health research is driven by widely recognised deficiencies of diagnostic categories alone for capturing mental disorders or providing adequate classes with which to cluster groups of patients for research or intervention. This is compounded by the lack of an instrument to capture symptomatology, as most research instruments would be considered overly cumbersome for routine clinical application outside specialist services. Furthermore, even if a fully structured instrument was identified as acceptable for use in initial assessment, obtaining real-time repeated measurements would present even more substantial challenges. The situation currently in mental health EHRs is that symptom profiles have been 'invisible' when it comes to deriving data for research, service development or clinical audit. Given that they are key determinants of interventions received and outcomes experienced, this has been a major deficiency. We therefore hope that the outputs of this project will offer the tools/techniques to use the large amounts of SMI symptomatology data contained within EHR systems, and provide new insight into the value of using SMI symptoms as predictors of a range of outcome measures. Although we did not seek to extend our analyses beyond simple descriptions of distributions, these strongly indicate that symptoms cross diagnostic groupings—for example, indicating that affective symptoms were not restricted to bipolar disorder. This is consistent with other reported findings from CRIS on mood instability which also cut across 'affective' and 'non-affective' psychosis[37] and which suggests that symptom dimensions rather than traditional diagnostic groupings may be a more valid approach to investigating aetiology and outcome in psychosis.

Contributors RGJ wrote the paper and analysed the data. RS and RP reviewed the manuscript, assisted with annotation and provided the clinical insight. AR and GG provided additional technical support. MB, NJ and AK provided annotation support. RJD provided additional supervisory guidance.

Funding All authors are funded by the National Institute of Health Research (NIHR) Biomedical Research Centre and Dementia Biomedical Research Unit at South London and Maudsley NHS Foundation Trust and King's College London. RP has received support from a UK Medical Research Council (MRC) Clinical Research Training Fellowship (MR/K002813/1) and a Starter Grant from the Academy of Medical Sciences.

Competing interests RJ, HS and RS have received research funding from Roche, Pfizer, J&J and Lundbeck.

Ethics approval Oxford REC C.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012;13:395–405.
2. Lin J, Jiao T, Biskupiak JE, *et al*. Application of electronic medical record data for health outcomes research: a review of recent literature. *Expert Rev Pharmacoecon Outcomes Res* 2013;13:191–200.
3. Shivade C, Raghavan P, Fosler-Lussier E, *et al*. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* 2014;21:221–30.
4. Tao C, Jiang G, Oniki TA, *et al*. A semantic-web oriented representation of the clinical element model for secondary use of

electronic health records data. *J Am Med Inform Assoc* 2013;20:554–62.
5. Rusanov A, Weiskopf NG, Wang S, *et al*. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. *BMC Med Inform Decis Mak* 2014;14:51.
6. Adam D. Mental health: on the spectrum. *Nature* 2013;496:416–18.
7. Chmielewski M, Bagby RM, Markon K, *et al*. Openness to experience, intellect, schizotypal personality disorder, and psychoticism: resolving the controversy. *J Personal Disord* 2014;28:483–99.
8. Van Os J, Marcelis M, Sham P, *et al*. Psychopathological syndromes and familial morbid risk of psychosis. *Br J Psychiatry* 1997;170:241–6.
9. Insel T, Cuthbert B, Garvey M, *et al*. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am J Psychiatry* 2010;167:748–51.
10. Andreasen NC. *Scale for the assessment of negative symptoms*. Iowa City: University of Iowa Press, 1983.
11. Kirkpatrick B, Strauss GP, Nguyen L, *et al*. The brief negative symptom scale: psychometric properties. *Schizophr Bull* 2010;37:300–5.
12. Kring AM, Gur RE, Blanchard JJ, *et al*. The Clinical Assessment Interview for Negative Symptoms (CAINS): final development and validation. *Am J Psychiatry* 2013;170:165.
13. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr Bull* 1987;13:261–76.
14. Axelrod BN, Goldman RS, Alphs LD. Validation of the 16-item Negative Symptom Assessment. *J Psychiatr Res* 1993;27:253–8.
15. Hardoon S, Hayes JF, Blackburn R, *et al*. Recording of severe mental illness in United Kingdom primary care, 2000–2010. *PLoS ONE* 2013;8:e82365.
16. Hersh WR. Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. *Am J Manag Care* 2007;13:277–8.
17. Embi PJ, Jain A, Clark J, *et al*. Development of an electronic health record-based Clinical Trial Alert system to enhance recruitment at the point of care. *AMIA Annu Symp Proc AMIA Symp AMIA Symp* 2005:231–5.
18. Embi PJ, Jain A, Clark J, *et al*. Effect of a clinical trial alert system on physician participation in trial recruitment. *Arch Intern Med* 2005;165:2272–7.
19. Antolík J. Automatic annotation of medical records. *Stud Health Technol Inform* 2005;116:817–22.
20. Nikiforou A, Ponirou P, Diomidous M. Medical data analysis and coding using natural language processing techniques in order to derive structured data information. *Stud Health Technol Inform* 2013;190:53–5.
21. Chapman WW, Nadkarni PM, Hirschman L, *et al*. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc* 2011;18:540–3.
22. Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17:507–13.
23. Tseytlin E, Mitchell K, Legowski E, *et al*. NOBLE—Flexible concept recognition for large-scale biomedical natural language processing. *BMC Bioinformatics* 2016;17:32.
24. Friedman C, Shagina L, Lussier Y, *et al*. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc* 2004;11:392–402.
25. Stewart R, Soremekun M, Perera G, *et al*. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. *BMC Psychiatry* 2009;9:51.
26. Perera G, Broadbent M, Callard F, *et al*. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. *BMJ Open* 2016;6:e008721.
27. Lukasiewicz M, Gerard S, Besnard A, *et al*. Young Mania Rating Scale: how to interpret the numbers? Determination of a severity threshold and of the minimal clinically significant difference in the EMBLEM cohort: YMRS severity cutoff. *Int J Methods Psychiatr Res* 2013;22:46–58.
28. Demjaha A, Morgan K, Morgan C, *et al*. Combining dimensional and categorical representation of psychosis: the way forward for DSM-V and ICD-11? *Psychol Med* 2009;39:1943.
29. Cuesta MJ, Peralta V. Integrating psychopathological dimensions in functional psychoses: a hierarchical approach. *Schizophr Res* 2001;52:215–29.
30. Jackson MSc RG, Ball M, Patel R, *et al*. TextHunter—a user friendly tool for extracting generic concepts from free text in clinical research. *AMIA Annu Symp Proc AMIA Symp AMIA Symp* 2014;2014:729–38.
31. Harkema H, Dowling JN, Thornblade T, *et al*. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomed Inform* 2009;42:839–51.
32. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol* 2011;2:1–27.
33. Taylor CL, Stewart R, Ogden J, *et al*. The characteristics and health needs of pregnant women with schizophrenia compared with bipolar disorder and affective psychoses. *BMC Psychiatry* 2015;15:88.
34. Fok ML-Y, Stewart R, Hayes RD, *et al*. The impact of co-morbid personality disorder on use of psychiatric services and involuntary hospitalization in people with severe mental illness. *Soc Psychiatry Psychiatr Epidemiol* 2014;49:1631–40.
35. Roque FS, Jensen PB, Schmock H, *et al*. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput Biol* 2011;7:e1002141.
36. Castro VM, Minnier J, Murphy SN, *et al*. Validation of electronic health record phenotyping of bipolar disorder cases and controls. *Am J Psychiatry* 2015;172:363–72.
37. Patel R, Lloyd T, Jackson R, *et al*. Mood instability is a common feature of mental health disorders and is associated with poor clinical outcomes. *BMJ Open* 2015;5:e007504

4.4.1 Supplementary File 1

NLP symptom model performance against gold standards (P=precision, R=recall, 95% confidence intervals in parenthesis)

Symptom	Model	Training Instances	TP	TN	FP	FN	P %	R %	F1
Aggression	ConText + ML	318	130	41	14	9	90 (+/- 5)	94 (+/- 4)	0.92
	ConText		135	22	33	4	80 (+/- 7)	97 (+/- 3)	0.88
Agitation	ConText + ML	296	170	11	10	1	94 (+/- 3)	99 (+/- 1)	0.97
	ConText		168	14	7	3	96 (+/- 3)	98 (+/- 2)	0.97
Anhedonia	ConText + ML	369	71	55	3	13	96 (+/- 4)	85 (+/- 8)	0.90
	ConText		83	52	6	1	93 (+/- 5)	99 (+/- 2)	0.96
Apathy	ConText + ML	287	90	20	4	23	96 (+/- 4)	80 (+/- 7)	0.87
	ConText		109	3	21	4	84 (+/- 7)	96 (+/- 3)	0.90
Arousal	ConText + ML	298	139	20	7	31	95 (+/- 3)	82 (+/- 6)	0.88
	ConText		169	6	21	1	89 (+/- 5)	99 (+/- 1)	0.94
Blunted or flat affect	ConText + ML	503	37	164	2	29	95 (+/- 5)	56 (+/- 12)	0.70
	ConText		66	18	148	0	31 (+/- 11)	100 (+/-0)	0.47
Catalepsy[1]	~								
Catatonic syndrome	ConText + ML	683	118	10	9	9	93 (+/- 4)	93 (+/- 4)	0.93
	ConText		122	6	13	5	90 (+/- 5)	96 (+/- 3)	0.93
Circumstantial speech	ConText + ML	271	79	323	12	17	87 (+/- 7)	82 (+/- 8)	0.84
	ConText		96	23	312	0	24 (+/- 8)	100 (+/-)	0.38
Deficient abstract thinking	ConText + ML	597	50	101	3	53	94 (+/- 4)	49 (+/- 1)	0.64
	ConText		100	26	78	3	56 (+/- 1)	97 (+/- 3)	0.71
Delusions	ConText + ML	677	99	28	3	8	97 (+/- 3)	93 (+/- 5)	0.95
	ConText		105	24	7	2	94 (+/- 5)	98 (+/- 3)	0.96
Derailment of speech	ConText + ML	280	104	15	5	4	95 (+/- 4)	96 (+/- 4)	0.96
	ConText		108	9	11	0	91 (+/- 5)	100 (+/-)	0.95
Diminished eye contact	ConText + ML	661	84	296	18	23	82 (+/- 7)	79 (+/- 8)	0.80
	ConText		86	4	310	21	22 (+/- 8)	80 (+/- 8)	0.34
Disturbed sleep	ConText + ML	750	58	27	10	11	85 (+/- 8)	84 (+/- 9)	0.85
	ConText		64	10	27	5	70 (+/- 11)	93 (+/- 6)	0.80

Symptom	Model	Training Instances	TP	TN	FP	FN	P %	R %	F1
Echolalia	ConText + ML	475	74	10	6	0	93 (+/- 6)	100 (+/- 0)	0.96
	ConText		73	7	9	1	89 (+/- 7)	99 (+/- 3)	0.94
Echopraxia[1]	~								
Elation	ConText + ML	335	177	39	20	11	90 (+/- 4)	94 (+/- 3)	0.92
	ConText		188	33	26	0	88 (+/- 5)	100 (+/- 0)	0.94
Elevated mood	ConText + ML	947	97	59	16	31	86 (+/- 6)	76 (+/- 7)	0.80
	ConText		125	37	38	3	77 (+/- 7)	98 (+/- 3)	0.86
Emotional withdrawal	ConText + ML	574	74	197	11	37	87 (+/- 6)	67 (+/- 9)	0.76
	ConText		110	47	161	1	41 (+/- 9)	99 (+/- 2)	0.58
Euphoria	ConText + ML	288	102	26	11	20	90 (+/- 5)	84 (+/- 7)	0.87
	ConText		120	19	18	2	87 (+/- 6)	98 (+/- 2)	0.92
Flight of ideas	ConText + ML	273	104	26	6	1	95 (+/- 4)	99 (+/- 2)	0.97
	ConText		103	29	3	2	97 (+/- 3)	98 (+/- 3)	0.98
Formal thought disorder	ConText + ML	605	97	139	15	6	87 (+/- 7)	94 (+/- 5)	0.90
	ConText		101	141	13	2	89 (+/- 6)	98 (+/- 3)	0.93
Grandiosity	ConText + ML	381	179	34	21	2	90 (+/- 4)	99 (+/- 2)	0.94
	ConText		176	34	21	5	89 (+/- 4)	97 (+/- 2)	0.93
Hallucinations	ConText + ML	1013	108	42	6	3	95 (+/- 4)	97 (+/- 3)	0.96
	ConText		110	44	4	1	96 (+/- 3)	99 (+/- 2)	0.98
Hostility	ConText + ML	581	163	13	18	2	90 (+/- 5)	99 (+/- 2)	0.94
	ConText		162	13	18	3	90 (+/- 5)	98 (+/- 2)	0.94
Immobility	ConText + ML	718	30	19	7	6	81 (+/- 13)	83 (+/- 12)	0.82
	ConText		36	4	22	0	62 (+/- 16)	100 (+/- 0)	0.77
Insomnia	ConText + ML	234	68	2	9	6	88 (+/- 7)	92 (+/- 6)	0.90
	ConText		73	3	8	1	90 (+/- 7)	99 (+/- 3)	0.94
Irritability	ConText + ML	632	146	16	23	12	86 (+/- 5)	92 (+/- 4)	0.89
	ConText		158	15	24	0	87 (+/- 5)	100 (+/- 0)	0.93
Loosening of associations	ConText + ML	353	0	1145	0	141	0 (+/- 0)	0 (+/- 0)	0.00
	ConText		141	110	1035	0	12 (+/- 5)	100 (+/- 0)	0.21

Symptom	Model	Training Instances	TP	TN	FP	FN	P %	R %	F1
Loss of coherence	ConText + ML	601	171	25	31	35	85 (+/- 5)	83 (+/- 5)	0.84
	ConText		203	4	52	3	80 (+/- 6)	99 (+/- 2)	0.88
Low mood	ConText + ML	879	0	35	0	21	0 (+/- 0)	0 (+/- 0)	0.00
	ConText		21	2	33	0	39 (+/- 21)	100 (+/- 0)	0.56
Mannerisms	ConText + ML	689	52	73	17	13	75 (+/- 1)	80 (+/- 1)	0.78
	ConText		62	36	54	3	53 (+/- 12)	95 (+/- 5)	0.69
Mutism[2]	Keyword		93	1	9	4	91 (+/- 6)	96 (+/- 4)	0.93
Negative syndrome	ConText + ML	150	114	12	11	12	91 (+/- 5)	90 (+/- 5)	0.91
	ConText		125	7	16	1	89 (+/- 6)	99 (+/- 2)	0.94
Paranoia	ConText + ML	263	144	10	8	3	95 (+/- 4)	98 (+/- 2)	0.96
	ConText		144	13	5	3	97 (+/- 3)	98 (+/- 2)	0.97
Persecutory ideation	ConText + ML	297	153	14	7	2	96 (+/- 3)	99 (+/- 2)	0.97
	ConText		152	15	6	3	96 (+/- 3)	98 (+/- 2)	0.97
Perseverance	ConText + ML	728	45	111	3	14	94 (+/- 6)	76 (+/- 11)	0.84
	ConText		59	11	103	0	36 (+/- 12)	100 (+/-)	0.53
Poor motivation	ConText + ML	753	0	249	0	116	0 (+/- 0)	0 (+/- 0)	0.00
	ConText		105	12	237	11	31 (+/- 8)	91 (+/- 5)	0.46
Poor rapport	ConText + ML	803	50	493	2	56	96 (+/- 4)	47 (+/- 1)	0.63
	ConText		90	6	489	16	16 (+/- 7)	85 (+/- 7)	0.26
Posturing	ConText + ML	1709	94	312	9	39	91 (+/- 5)	71 (+/- 8)	0.80
	ConText		131	46	275	2	32 (+/- 8)	98 (+/- 2)	0.49
Poverty of speech	ConText + ML	373	172	54	17	18	91 (+/- 4)	91 (+/- 4)	0.91
	ConText		189	33	38	1	83 (+/- 5)	99 (+/- 1)	0.91
Poverty of thought	ConText + ML	368	97	7	2	8	98 (+/- 3)	92 (+/- 5)	0.95
	ConText		98	6	3	7	97 (+/- 3)	93 (+/- 5)	0.95
Pressured speech	ConText + ML	377	268	11	14	23	95 (+/- 2)	92 (+/- 3)	0.94
	ConText		285	8	17	6	94 (+/- 3)	98 (+/- 2)	0.96
Rigidity	ConText + ML	1331	43	90	11	14	80 (+/- 1)	75 (+/- 11)	0.77
	ConText		56	41	60	1	48 (+/- 13)	98 (+/- 3)	0.65
Social withdrawal	ConText + ML	307	137	33	24	18	85 (+/- 6)	88 (+/- 5)	0.87
	ConText		153	4	53	2	74 (+/- 7)	99 (+/- 2)	0.85

Symptom	Model	Training Instances	TP	TN	FP	FN	P %	R %	F1
Stereotypy	ConText + ML	1397	0	72	0	64	0 (+/- 0)	0 (+/- 0)	0.00
	ConText		63	23	49	1	56 (+/- 12)	98 (+/- 3)	0.72
Stupor	ConText + ML	567	84	21	12	11	88 (+/- 7)	88 (+/- 6)	0.88
	ConText		93	8	25	2	79 (+/- 8)	98 (+/- 3)	0.87
Tangential speech	ConText + ML	250	101	0	7	0	94 (+/- 5)	100 (+/- 0)	0.97
	ConText		99	2	5	2	95 (+/- 4)	98 (+/- 3)	0.97
Thought block	ConText + ML	377	85	32	9	41	90 (+/- 5)	67 (+/- 8)	0.77
	ConText		124	19	22	2	85 (+/- 6)	98 (+/- 2)	0.91
Waxy flexibility	ConText + ML	453	66	91	6	13	92 (+/- 6)	84 (+/- 8)	0.87
	ConText		76	51	46	3	62 (+/- 11)	96 (+/- 4)	0.76

1. All instances annotated. 2. Keyword only – no NLP applied

4.4.2 Additional discussion of CRIS-CODE results

Upon request from the thesis reviewers, below I provide an addendum to the discussion of the results of the CRIS-CODE paper.

The classification results of certain symptoms using both the naive ConText approach and the hybrid approach was observered to be poorer than average. These include loosening of associations (f1 of hybrid: 0.00, f1 of ConText: 0.21), blunted or flat affect (f1 of hybrid: 0.70, f1 of ConText: 0.47), deficient abstract thinking (f1 of hybrid: 0.64, f1 of ConText: 0.71), low mood (f1 of hybrid: 0.00, f1 of ConText: 0.56), poor rapport (f1 of hybrid: 0.63, f1 of ConText: 0.26), poor motivation (f1 of hybrid: 0.00, f1 of ConText: 0.46) and stereotypy (f1 of hybrid: 0.00, f1 of ConText: 0.72). Although an extensive investigation into the failure of my approaches to produce comparable results to other symptoms did not take place, I offer the following additional possible reasons for why these symptoms yielded poor results.

In the cases of loosening of associations, poor motivation, low mood and stereotypy, the ML approach failed catastrophically, and not a single true positive was correctly predicted. These are particularly curious results, as the classifier produced a result that was worse than chance. This seems to suggest that 1) there may be some flaw in the training data, such that no consistently predictive features could be obtained by the classifier, or 2) the classifier failed silently during training for an unknown technical reason (the Batch Learning plugin used to build these models in GATE has now been deprecated). Future work might attempt to retrain a classifier for these examples using a different implementation of SVMs and/or the external code used to create the feature matrix prior to training (for instance, the scikit-learn implementations in Python). If such work were to reveal that the classifier is still unable to produce a modest improvement in classification performance, the likely culprit would be an issue with the training data. Nevertheless, as the baseline classification performance from ConText in isolation was also not encouraging for these symptoms, it seems likely that they also suffer from disambiguation issues as alluded to in the original manuscript.

Deficient abstract thinking, poor rapport and blunted or flat affect all produced notably lower f1 scores than other symptoms in both the hybrid and ConText models. Since the ML classifier was able to produce better than random results, this suggests that a higher f1 might be obtained via increasing the volume of training data made available to them. As discussed above, such cases would also most likely benefit from offering more complex features. However, in light of recent, substantial progress made via neural network classification methodologies such ULMfit [232] and Bert [233], a more rational

approach to future work would rethink the entire TextHunter pipeline to make use of such advances.

4.5 Other projects using the TextHunter methodology

To date, CRIS-CODE is the largest annotation and model building project that has been undertaken with TextHunter. However, the TextHunter methodology has been successfully applied in a number of other initiatives. These include:

1. The occurence of mood instability in various mental health disorders [11]
2. An investigation of novel psychoactive substances in social media and electronic health records [234]
3. Cannabis use, hospitalisation and antipsychotic treatment failure in first episode psychosis [10]
4. Investigating clinical outcomes in psychotic disorders [235]
5. Violent topologies amongst women inpatients with severe mental illness [236]
6. Analysis of diagnoses extracted from a large mental health case register [237]
7. Negative symptoms in early onset psychosis and their association with antipsychotic treatment failure [238]

4.6 Conclusion

In collaboration with the University of Sheffield, I successfully identified and validated a methodology to support simple IE tasks in the context of negative symptomatology. With the development of the TextHunter software, I have further refined this process to make it viable for more general information extraction tasks. An important aspect of the TextHunter software has been to streamline this technique to enable a degree of self service for lay users of text analytics. This has culminated in the CRIS-CODE project. CRIS-CODE was largely successful in its ambition, providing detailed symptom profiles of the SLAM SMI population. However, the process of reviewing and annotating large volumes of clinical documents revealed some interesting observations regarding the diversity of language used by clinicians in their patient facing interviews. These observations are explored further in the next chapter. Finally, the popularity of the TextHunter methodology as demonstrated by CRIS-CODE and its uptake in a range of other projects is consistent

with the argument I laid out in chapter 3, regarding that simple NLP methodologies packaged in a streamlined fashion is sufficient to address a small subset of real world clinical data challenges.

Chapter 5

Knowledge Discovery for Deep Phenotyping Serious Mental Illness from Electronic Health Records

5.1 Overview

In the previous chapter, I described a large scale project to convert unstructured textual data into structured symptomatology data. Once structured, such data can be represented in a vast number of ways. However, in the course of the annotation process of work, it became apparent that a complex array of language constructs were in use within the CRIS corpus, that might not necessarily reflect the top-down definitions of SMI symptomatology we originally produced. To explore this further, I present the following article (author contributions are listed in the paper):

RESEARCH ARTICLE

REVISED Knowledge discovery for Deep Phenotyping serious mental illness from Electronic Mental Health records [version 2; referees: 2 approved]

Richard Jackson [1], Rashmi Patel [1,2], Sumithra Velupillai[1,3], George Gkotsis[1], David Hoyle [4], Robert Stewart [1,2]

[1]Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, SE5 8AF, UK
[2]South London and Maudsley NHS Foundation Trust, London, SE5 8AZ, UK
[3]School of Computer Science and Communication, TH Royal Institute of Technology, Stockholm, SE-100 44, Sweden
[4]Independent Researcher, Manchester, UK

v2 **First published:** 21 Feb 2018, **7**:210 (doi: 10.12688/f1000research.13830.1)
Latest published: 08 May 2018, **7**:210 (doi: 10.12688/f1000research.13830.2)

Abstract

Background: Deep Phenotyping is the precise and comprehensive analysis of phenotypic features in which the individual components of the phenotype are observed and described. In UK mental health clinical practice, most clinically relevant information is recorded as free text in the Electronic Health Record, and offers a granularity of information beyond what is expressed in most medical knowledge bases. The SNOMED CT nomenclature potentially offers the means to model such information at scale, yet given a sufficiently large body of clinical text collected over many years, it is difficult to identify the language that clinicians favour to express concepts.
Methods: By utilising a large corpus of healthcare data, we sought to make use of semantic modelling and clustering techniques to represent the relationship between the clinical vocabulary of internationally recognised SMI symptoms and the preferred language used by clinicians within a care setting. We explore how such models can be used for discovering novel vocabulary relevant to the task of phenotyping Serious Mental Illness (SMI) with only a small amount of prior knowledge.
Results: 20 403 terms were derived and curated via a two stage methodology. The list was reduced to 557 putative concepts based on eliminating redundant information content. These were then organised into 9 distinct categories pertaining to different aspects of psychiatric assessment. 235 concepts were found to be expressions of putative clinical significance. Of these, 53 were identified having novel synonymy with existing SNOMED CT concepts. 106 had no mapping to SNOMED CT.
Conclusions: We demonstrate a scalable approach to discovering new concepts of SMI symptomatology based on real-world clinical observation. Such approaches may offer the opportunity to consider broader manifestations of SMI symptomatology than is typically assessed via current diagnostic frameworks, and create the potential for enhancing nomenclatures such as SNOMED CT based on real-world expressions.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
REVISED version 2 published 08 May 2018	✓ report	✓ report
version 1 published 21 Feb 2018	? report	? report

1 **Julian Hong**, Duke University School of Medicine, USA
Jessica Tenenbaum, Duke University School of Medicine, USA

2 **Karin Verspoor**, The University of Melbourne, Australia
The University of Melbourne, Australia

Discuss this article

Comments (0)

Keywords
word2vec, natural language processing, serious mental illness, electronic health records, schizophrenia

Corresponding author: Robert Stewart (robert.stewart@kcl.ac.uk)

Author roles: Jackson R: Conceptualization, Formal Analysis, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing – Original Draft Preparation; **Patel R**: Data Curation, Validation, Writing – Review & Editing; **Velupillai S**: Methodology, Writing – Review & Editing; **Gkotsis G**: Methodology, Writing – Review & Editing; **Hoyle D**: Methodology, Supervision, Writing – Review & Editing; **Stewart R**: Data Curation, Funding Acquisition, Project Administration, Resources, Supervision, Validation, Writing – Review & Editing

Competing interests: RJ and RS have received research funding from Roche, Pfizer, J&J and Lundbeck.

How to cite this article: Jackson R, Patel R, Velupillai S *et al.* **Knowledge discovery for Deep Phenotyping serious mental illness from Electronic Mental Health records [version 2; referees: 2 approved]** *F1000Research* 2018, **7**:210 (doi: 10.12688/f1000research.13830.2)

Grant information: This paper represents independent research funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. RP has received support from a Medical Research Council (MRC) Health Data Research UK Fellowship and a Starter Grant for Clinical Lecturers (SGL015/1020) supported by the Academy of Medical Sciences, The Wellcome Trust, MRC, British Heart Foundation, Arthritis Research UK, the Royal College of Physicians and Diabetes UK. SV is supported by the Swedish Research Council (2015-00359), Marie Sklodowska Curie Actions, Cofund, Project INCA 600398
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 21 Feb 2018, **7**:210 (doi: 10.12688/f1000research.13830.1)

REVISED **Amendments from Version 1**

This revision includes amendments that we hope address the issues raised by the peer review process. A response to each comment can be found in the 'response to reviewer' section that accompanies the article, but the changes can be summarised as follows:

1. Improvements to the clarity of the methods section, addressing some comprehension issues that were raised such as consistency of terminology and the description of techniques employed
2. An expanded rationale for several decisions that were made in the development of the approach, against alternatives that were available
3. The citation of additional relevant literature for this domain, such as work on automated term recognition and existing work on symptom grouping
4. Some additional results regarding the counts of unigrams, bigrams and trigrams
5. A reference to a publicly available code repository that demonstrates the approach (since sharing the underlying data is not possible)
6. Several minor grammatical errors

We offer our gratitude to both sets of reviewers for their time and valuable assistance.

See referee reports

Introduction

The dramatic decrease of genetic sequencing costs, coupled with the growth of our understanding of the molecular basis of diseases, has led to the identification of increasingly granular subsets of disease populations that were once thought of as homogenous groups. As of 2010, the molecular basis for nearly 4 000 Mendelian disorders has been discovered[1], subsequently leading to the development of around 2 000 clinical genetic tests[2]. The resulting 'precision medicine' paradigm has been touted as the logical evolution of evidence-based medicine.

Precision medicine has arisen in response to the fact that the real-world application of many treatments have a lower efficacy and a differential safety profile compared to clinical trials, most likely due to genetic and environmental differences in the disease population. Precision medicine seeks to obtain deeper genotypic and phenotypic knowledge of the disease population, in order to offer tailored care plans with evidence-based outcomes. Amongst the challenges presented by precision medicine is the requirement to obtain highly granular phenotypic knowledge that can adequately explain the variable manifestation of disease.

To realise the ambitions of precision medicine, large amounts of phenotypic data are required to provide sufficient statistical power in tightly defined patient cohorts (so called 'Deep Phenotyping'[3]). Historical clinical data mined from Electronic Health Record (EHR) systems are frequently employed to meet the related use case of observational epidemiology. As such, EHRs are often posited as the means to provide extensive phenotypic information with a relatively low cost of collection[4,5].

In order to standardise knowledge representation of clinically relevant entities and the relationships between them, phenotyping from EHRs often employs curated terminology systems, most commonly SNOMED CT. The use of such resources creates a common domain language in the clinical setting, theoretically allowing an unambiguous interpretation of events to be shared within and between healthcare organisations. The anticipated value of such a capability has prompted the UK National Information Board to recommend the adoption of SNOMED CT across all care settings by 2020[6]. However, the task of representing the sprawling and ever-changing landscape of healthcare in such a fashion has proven complex[7–10]. Although a complete description of the structure and challenges of SNOMED CT are beyond the scope of this paper, we describe how aspects of these problems manifest themselves in accordance with the task of phenotyping serious mental illness (SMI) from a real-world EHR system.

Phenotyping SMI

The quest for empirically validated criteria for assessing the symptomatology of mental illness has been a long term goal of evidence-based psychiatry. SMI is a commonly used umbrella term to denote the controversial diagnoses of schizophrenia (encoded in SNOMED as SCTID: 58214004), bipolar disorder (SCTID: 13746004), and schizoaffective disorder (SCTID: 68890003). While field trials of DSM-5 have revealed promising progress in reliably delineating these three conditions in clinical assessment[11], such diagnostic entities continue to have low clinical utility[12–14]. Recent evidence from genome-wide association studies appears to suggest that such disorders share common genetic loci, further countering the argument that SMI can be classified into discrete, high level diagnostic units[15]. In terms of clinical practice, the presenting symptomatology of SMI is usually the basis for treatment. This is often characterised by abnormalities in various mental processes, which are in turn categorised according to broad groupings of clinically observable behaviours. For instance, 'positive symptoms' refer to the presence of behaviours not seen in unaffected individuals, such as hallucinations, delusional thinking and disorganised speech. Conversely, 'negative symptoms', such as poverty of speech and social withdrawal refer to the absence of normal behaviours. Such symptomatology assessments are organised via an appropriate framework such as Postive and Negative Symptom Scale[16] (PANSS) or Brief Negative Symptom Scale[17]. Accordingly, SNOMED CT includes coverage for many of these symptoms, generally within the 'Behaviour finding' branch (SCTID: 844005).

A qualifying factor regarding the adoption of SNOMED amongst SMI specialists might therefore require that the list of clinical 'finding' entities in SNOMED are sufficiently expansive and diverse to represent their own experiences during patient interactions. Specifically, this may manifest as two key challenges for terminology developers.

First, insight must be obtained regarding real-world language usage such that universally understood medical concepts, encompassing hypernomy, synonymy and hyponomy. Similarly, the abundant use of acronyms in the medical domain means that a

large percentage of acronyms to have two or more meanings[18], creating word sense disambiguation problems. As such, significant efforts have arisen to supplement these types of knowledge bases with appropriate real-world synonym usage extracted from EHR datasets[19]. The problem may be considered analogous to difficulties in the recognition, classification and mapping of technical terminology variants throughout the biomedical literature, which is known to be an impediment to the construction of knowledge representation systems (see 20 for a review).

Second, if there is controversy over international consensus in a particular area of medicine, the use of 'global' perspectives may not be sufficient to meet local reporting/investigatory requirements. Such issues are particularly pertinent in mental health where many diseases defy precise definition and biomarker development has yielded few successes[21]. More generally, all medical knowledge bases are incomplete to one degree or another. The opportunity to utilise large amounts of EHR data to discover novel observations and relationships arising from real-world clinical practise must not be overlooked.

Given a sufficiently large corpus of documents, typically written by hundreds of clinical staff over several years, it is often difficult to track the evolution of vocabulary used within the local EHR setting to describe potentially important clinical constructs. In previous work, we describe our attempts to extract fifty well known SMI symptomatology concepts from a large electronic mental health database resource[22], based upon the contents of such frameworks. During the course of manually reviewing clinical text, we made two subjective observations of the documentation resulting from clinician/patient interactions:

- The tendency of clinicians to use non-technical vocabulary in describing their observations
- The occasional appearance of highly detailed, novel observations that do not readily fit into known symptomatology frameworks

Such observations may feasibly have clinical relevance, for example, as non-specific symptomatology prodromes[23]. On the basis that the modelling of SMI for precision medicine approaches require the full dimensionality of the disease to be considered, we sought to explore these observations further.

In this study, we present our efforts to utilise *a priori* knowledge discovery methods to identify preferences in real-world language usage that reflect clinically relevant SMI symptomatology within the context of a large mental healthcare provider. We contrast and compare these patterns with a modern version of the UK SNOMED CT (v1.33.2), and suggest how such approaches may offer novel and/or more granular symptom expressions from patient/clinician interactions when used to supplement resources such as SNOMED CT, potentially offering alternatives to classify psychiatric disorders with finer resolution and greater real-world validity.

Methods

Our general approach for SMI knowledge discovery is composed of several discrete steps. An overview of the workflow is given in Figure 1.

Corpus creation from the Clinical Record Interactive Search

The South London and Maudsley NHS Foundation Trust (SLaM) provides mental health services to 1.2 million residents over four south London boroughs (Lambeth, Southwark, Lewisham and Croydon). Since 2007, the Clinical Record Interactive Search (CRIS)[24] infrastructure programme has been operating to offer a pseudonymised and de-identified

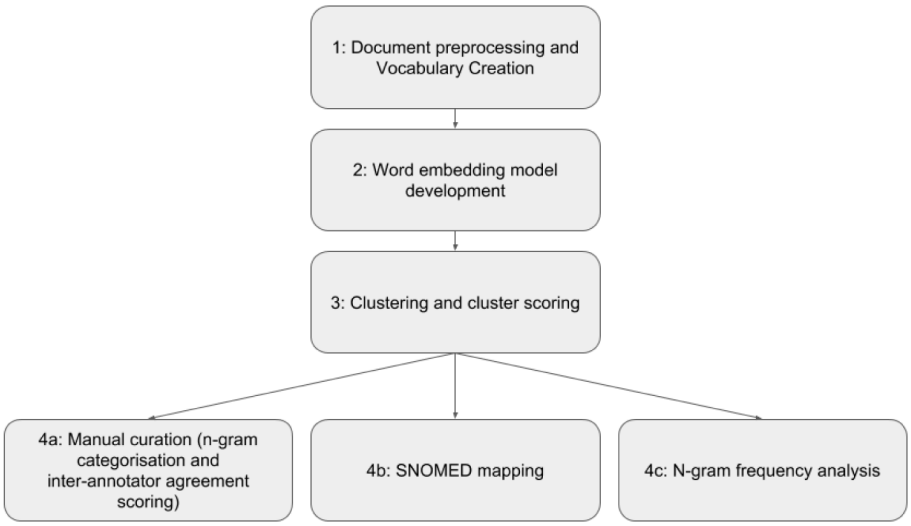


Figure 1. Overview of project workflow.

research database of SLaM's EHR system. As the CRIS resource received ethical approval as a pseudonymised and de-identified data source by Oxford Research Ethics Committee (reference 08/H0606/71+5), patient consent was not required for this study.

11 745 094 clinical documents were collected from the CRIS database from the period 01/01/2007 - 27/10/2016 on the basis that the 20 472 associated patients were assigned an SMI ICD10 code of F20, F25, F30 or F31 at some point during their care, in accordance with current clinical practice.

Pre-processing and vocabulary creation

Sentences and tokens were extracted from each document using the English Punkt tokeniser from the NLTK 3.0 suite[25]. Each token was converted to lower case. A vocabulary was then constructed of all 1-gram types in the corpus, supplemented with frequently occuring bi-grams and tri-grams using the Gensim[26] suite and the sampling method proposed by Mikolov *et al.*[27]. Bi-grams and tri-grams with a minimum frequency of 10 occurrences in the entire corpus were retained, to give a total vocabulary size of 896 195 terms (617 095 unigrams, 277 490 bigrams, 303 trigrams and 1 307 non-word entities). No further assumptions about the structure of the data, such as the need for stemming/lemmatisation, were made.

Building a word embedding model

The distributional hypothesis was first explored by Harris[28], which proposed that, given a sufficiently large body of text, linguistic units that co-occur in the same context are likely to have a semantically related meaning. Modelling the distribution of such units may therefore have value for a wide range of natural language processing applications. Models of distributional semantics, including word embeddings, are techniques that aim to derive models of semantically similar units in a corpus of text by co-locating them in vector space. In recent years, the use of the Continuous Bag-of-Words (CBOW) model proposed by Mikolov *et al.*[29] has risen to prominence, owing to its ability to accurately capture semantic relationships whilst scaling to large corpora of text[27]. Recently, the CBOW model has been used to identify the semantic similarities between single word entities in biomedical literature and clinical text[30], suggesting that biomedical literature may serve as a useful proxy for clinical text, for tasks such as synonym identification and word sense disambiguation tasks under limited conditions[30].

A full description of the CBOW architecture is discussed in 31. For brevity, we describe only the key features used in our work here. The purpose of the architecture is to 'learn' in an unsupervised manner, a representation of the semantics of different terms, given an input set of documents. CBOW might be described as a simple feed forward neural network consisting of three layers. An input layer X composed of o nodes (where o is the number of unique terms in a corpus produced from our above described pre-processing), a hidden layer H of a user defined size n (usually between 100 and 300), and an output layer Y that is also composed of o nodes. Every node in X is connected to every node in H, and every node in H is connected to every node in Y . Between each of the layers is a matrix of weight values; for the X and H layer, an 'input' matrix of dimensions $o \times n$ (hereafter denoted W); and between the H and the Y layer, an 'output' matrix of dimensions $n \times o$ (denoted W'). The output of training the neural network is to produce weights in each of these matrices. The weights learnt in the W matrix might be intuitively described as the semantic relationships between each term in the vocabulary as represented in vector space, with semantically similar words located in closer proximity to each other. Weights in the W' matrix represent the predictive model from the H to the Y layer. A training instance is composed of a group of terms, known as a context. A context can be composed of natural language structures, such as sentences in a document, or more complex arrangements, such as a sliding window of terms (usually between 5 and 10) that move over each token in a document (potentially ignoring natural grammatical structures). For a given input term, the input into the nodes on the hidden layer is the product of each vector index in matrix W corresponding to each context word and the average vector. From the H to the Y layer, it is then possible to score each term using the W' matrix, from which a posterior probability is obtained for each word in the vocabulary using the softmax function. The weights in each matrix are then updated using computationally efficient hierarchical softmax or negative sampling approaches. Once training is complete, the semantic similarity of terms is often measured via their cosine distance between vectors in the W matrix.

Using the Gensim implementation of CBOW and our previously constructed vocabulary, we trained a word embedding model of $n = 100$ over our SMI corpus to produce a vector space representation of our clinical vocabulary. Due to patient confidentiality, offline access to records was not feasible and so only a limited number of epochs of training could be performed. However, due to the relatively narrow/controlled vocabulary employed in clinical records (compared to normal speech/text) the range of possible input vectors was narrower than might otherwise be expected, and even a single epoch of training appeared to yield meaningful clusters that could be identified with SMI. As we were primarily intending to identify initial clusters for validation by clinical experts it was felt that single epoch of training, over the 20M clinical records available, was sufficient.

Vocabulary clustering and cluster scoring

The task of clustering seeks to group similar dataset objects together in meaningful ways. In unsupervised clustering, the definition of 'meaningfulness' is often subjectively defined by the human observers. In our task, we sought to identify clusters of terms derived from our word embedding model that represent semantically linked components of our clinical vocabulary, based on the theory that our word embedding model would cause related symptom concepts to appear close to each other within the vector space.

A particular challenge in the development of clustering algorithms is achieving scalability to large datasets. Since many clustering algorithms make use of the pairwise distance between n samples (or terms, in our case), the memory requirements of such algorithms tend to run in the order of n^2. One such algorithm that does not suffer from this limitation is k-means clustering. k-means clustering is a partitional clustering algorithm that seeks to assign n samples into a user defined k clusters by minimising the squared error between each centroid of a cluster and its surrounding points. A global (although not necessarily optimal) solution is derived when the algorithm has minimised the sum of squared errors across all k clusters, subject to some improvement threshold or other stopping criteria. For all experiments, we used the k-means++ implementation from the Scikit-Learn framework[32] with 8 runs each time, to control against centroids emerging in local minima.

The key parameter for k-means clustering is the selection of k. While techniques exist for estimating an appropriate value, such as silhouette analysis and the 'elbow method'[33], these utilise pairwise distances between samples, creating substantial technical limitations for large matrices in terms of memory usage. To overcome this, we opted for a memory efficient version of the elbow method, involving plotting the minimum centroid distance for different values of k. The intuition behind this approach is that every increase in k is likely to result in a smaller minimum centroid distance in vector space (subject to a random seed for the algorithm). As k increases, genuine clusters should be separated by a steady decline in minimum centroid distance. However, when the slope of the decline flattens out (i.e. the 'elbow' of the curve), assignment of samples to new clusters is likely to be random).

With the data clustered, we sought to identify one or more clusters of interest for further examination. To this end, we devised a simple 'relevance' cluster scoring approach based upon prior knowledge of common SMI symptom concepts. The intuition behind our approach is that the training of the Word2Vec model will cause terms that represent 'known' concepts of SMI symptomatology to colocate in close proximity to each other in vector space, and the clustering approach will place them in the same cluster, along with other terms that theoretically relate to these SMI symptomatology concepts. The additional contents of this cluster may therefore hold terms that represent concepts of SMI symptomatology undefined by our team, but in natural use by the wider clinical staff of the SLaM Trust during the course of their duties. By identifying the richest cluster(s) in terms of the known SMI symptomatology lexicon, we sought to drastically reduce the search space of terms in the corpus to carry forward for human assessment.

We selected 38 internationally recognised symptom concepts of SMI based upon their expression in SMI frameworks and on their specificity in clinical use (Table 1), to form the basis of our scoring algorithm. For instance, we did not select 'loosening of associations', due to the different word sense that the word 'associations' appears in, such as 'housing associations', and organisational references such as 'Stroke Association'.

Table 1. Known symptomatology concepts and Prior Concept vocabulary matching sequences used for cluster scoring. An underscore represents a bigram match.

SMI symptom	Prior Concept matching character sequence
aggression	aggress
agitation	agitat
anhedonia	anhedon
apathy	apath
affect	affect
catalepsy	catalep
catatonic	cataton
circumstantial	circumstant
concrete	concrete
delusional	delusion
derailment	derail
eye contact	eye_contact
echolalia	echola
echopraxia	echopra
elation	elat
euphoria	euphor
flight of ideas	foi
thought disorder	thought_disorder
grandiosity	grandios
hallucinations	hallucinat
hostility	hostil
immobility	immobil
insomnia	insomn
irritability	irritab
coherence	coheren
mannerisms	mannerism
mutism	mute
paranoia	paranoi
persecution	persecut
motivation	motivat
rapport	rapport
posturing	postur
rigidity	rigid
stereotypy	stereotyp
stupor	stupor
tangential	tangenti
thought block	thought_block
waxy	waxy

Rather, we chose symptoms such as 'aggression', 'apathy' and 'agitation', which are less likely to have different word sense interpretations in the context of SMI clinical documents.

For each of the 38 concepts, we produced a set of terms constituting stems and appropriate synonyms/acronyms as described

in Table 1, in order to produce a set of character sequences representing existing domain knowledge, or 'prior concepts' (hereafter, termed PCs) that could be matched against each term in each cluster via regular expressions. With this matching criterion, we scored each cluster based on the number of hits to derive a cluster/PC count matrix x where $x_{i,j}$ represents the count of the ith PC in the jth cluster. For example, a cluster containing the 1-gram 'insomnia' and 'insomniac' would receive a count of two for the 'insomni' PC. For each PC, we then calculated a vector of the minimum count per concept across all clusters:

$$u_i = min_{j \in J} x_{ij},\ i = 1, \ldots, m. \quad (1)$$

where m is 38 (denoting the number of PCs we describe in Table 1). Similarly, we generated a vector of maximum count per PC across all clusters:

$$v_i = max_{j \in J} x_{ij},\ i = 1, \ldots, m. \quad (2)$$

to enable us to rescale the value of each PC/cluster count to between 0 and 1 into a matrix x':

$$x'_{ij} = \frac{x_{i,j} - u_i}{v_i - u_i} \quad (3)$$

The purpose of rescaling in such a way was to prevent overrepresented PCs unduly influencing the overall result (for instance, a PC with many hits in a cluster would unduly bias the score towards that concept, whereas we sought a scoring mechanism that would weigh all input PCs equally, regardless of their frequency).

Finally, we summed all rescaled PC counts per cluster, and divided by the total cluster size to provide a score per cluster z representing the value of the:

$$z_j = \frac{\sum_{i=1}^{m} x'_{ij}}{s_j} \quad (4)$$

where s is a vector of the total count of terms in each cluster. The purpose of dividing by cluster size was to prevent the tendency of larger clusters to score higher on account of their size.

To select clusters for further investigation, the robust median absolute deviation (MAD) statistic was chosen (the distribution of our cluster scores was non-normal). This precipitated clusters that were the most valuable, in terms of the breadth of PC concept hits they contain. We adopted a conservative approach to cluster selection by choosing clusters that scored at least six MAD above the median score for further processing, which is approximately equivalent to four standard deviations for a normally distributed dataset.

We provide a worked example of this technique in the code repository that accompanies this paper, using publically available data.

Expert curation of symptom concepts, frequency analysis and SNOMED CT mapping

The contents of the top scoring clusters underwent a two stage curation process. The first stage was performed by an informatician, and involved several simple string processing tasks to filter out uninteresting terms. Such processes included removal of terms that contained tokenisation failures (for example, single character non-word tokens such as 'y', 'p') and other constructs that had low information content, such as terms composed of stop words. A final manual check followed to reduce the annotator burden required by the clinical team.

The second, more important, stage was composed of independent annotation of the curated concept list by two psychiatrists, to identify likely synonyms and new symptomatology based on their clinical experience. Each concept was assigned to one of the below 8 'substantive' categories, or a 9th 'other' category. The categories were derived from 34, and the experience of the team Clinical Psychiatrists.

Appearance/Behaviour Implying a real-time description of the way a patient appears or behaves (including their interaction with the recording clinician)

Speech Anything implying a description of any vocalisation (i.e. theoretically a subset of behaviour but restricted to vocalisations)

Affect/Mood Implying clinician-observed mood/emotional state (i.e. theoretically a subset of appearance but restricted to observed emotion), or implying self-reported mood/emotional state (i.e. has to imply a description that a patient would make of their own mood; theoretically a subset of thought)

Thought Implying any other thought content

Perception Implying any described perception

Cognition Implying anything relating to the patient's cognitive function

Insight Implying anything relating to insight (awareness of health state)

Personality Anything implying a personality trait or attitude (i.e. something more long-standing than an observed behaviour at interview)

Other A mixed bag of definable terms that do not fit into the above. Common examples included anything implying information that will have been collected as part of a patient's history, often of behaviours that would have to have been reported as occurring in the past and cannot have been observed at interview, but also which cannot be termed a personality trait. Alternatively, anything where insufficient context was available to make a decision

Inter annotator agreement (IAA) was measured with the Cohen's Kappa agreement statistic[35].

To explore the frequency of both our prior symptomatology concepts and the newly curated ones in our symptom clusters, we counted the number of unique patient records and the number of unique documents in which the stems of each term appeared. To protect patient anonymity, we discarded any concept that appeared in ten or fewer unique patient records. Finally, we mapped the remaining concepts to SNOMED CT, UK version v1.33.2, using the following method. First, the root mopheme of each concept was matched to a relevant finding, observable entity or disorder type in SNOMED CT. If a match could not be found, SNOMED CT was explored for potential synonymy, or other partial match. If a clear synonym could not be found, we classified the concept as novel.

Results

Word embedding model training

Processing the corpus of SMI clinical documents took approximately 100 hours on an 8-core commodity hardware server. Documents were fed sequentially from an SQL Server 2008 database operating as a shared resource, with an additional overhead likely resulting from network latency.

Parameter selection for *k*-means clustering

Figure 2 shows a scatterplot of variable values of k and the resulting minimum centroid distance. This suggests a k value of around 50–75 may be optimal for our data. On this basis, we chose a k value of 75.

Cluster scoring

The application of our relevancy scoring algorithm to the 75 derived clusters resulted in a median score was 0.000229 and a MAD of 0.000277, and is visualised in Figure 3.

Three clusters emerged with a score at least six MADs outside of the median cluster score: No. 52 (score: 0.002883), containing 6 665 terms, No. 69, containing 9 314 (score: 0.002282) terms and No. 49 (score: 0.001940), containing 4 424 terms. Taken together, these three clusters contained a total of 20 403 terms.

Expert curation of symptom concepts, frequency analysis and SNOMED CT mapping

The combined 20 403 terms were taken forward for curation as described above. The first phase of curation reduced the list to 519 putative concepts. The majority of eliminated terms were morphological variations, misspellings and tokenisation anomalies of singular concepts. For instance, 84 variations were detected for the stem 'irrit*' (as in 'irritable'). Other terms were removed because insufficient context was available for a reasonable clinical interpretation, such as 'fundamentally unchanged', 'amusing' and 'formally tested'. Finally, terms that appeared to have no relevance to symptomatology at all were removed, such as dates and clinician names.

Expert curation by two psychiatrists of the 557 concepts (519 discovered concepts and 38 prior concepts) produced a Cohen's Kappa agreement score of 0.45, where 337 concepts were assigned to one of our 9 categories independently by expert psychiatric curation. Of the 337 concepts, 235 were assigned to a substantive category (i.e. not the indeterminate 'other' group). Table 2 shows the number of terms per category where agreement was reached.

Supplementary File 1 is a CSV table of all 557 terms. In addition to the term itself, the table contains the following information; the counts of the unique patient records of our 20 472 patient SMI cohort in which the term was detected; the counts of the unique documents of the 11 745 094 clinical

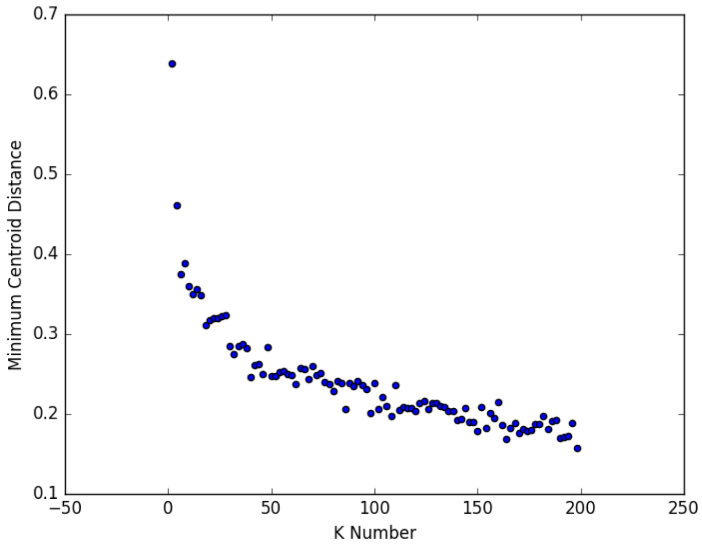


Figure 2. Selecting K for K-means++.

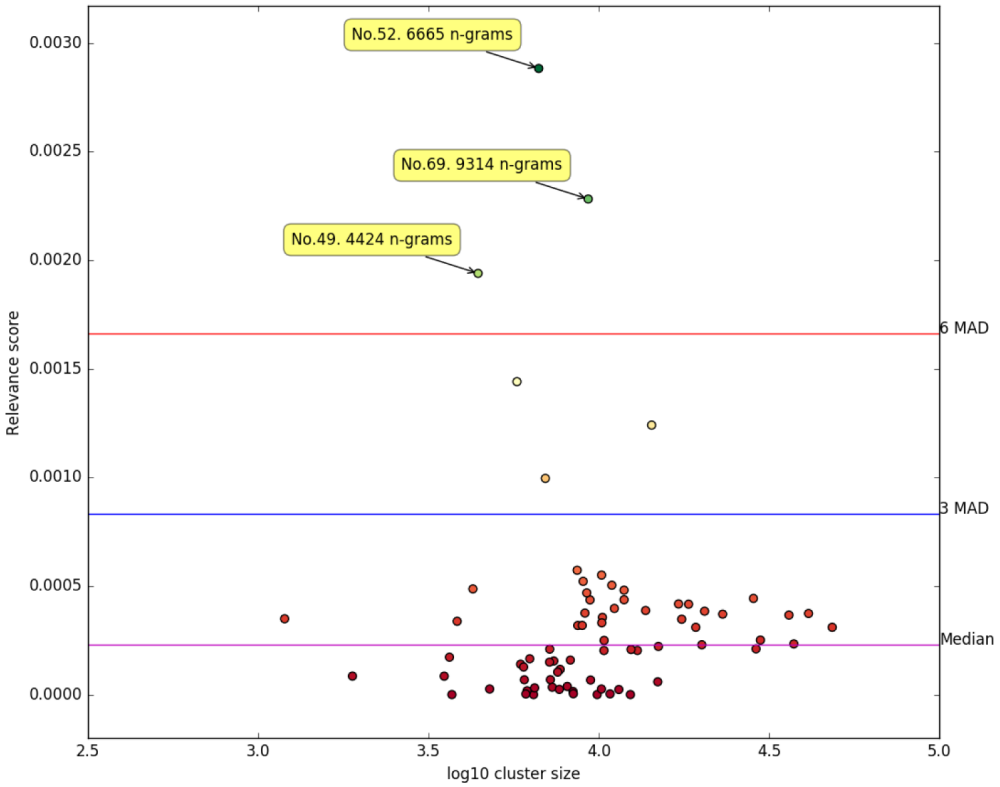


Figure 3. Scoring of clusters according to known symptomatology content. Each dot represents a unique cluster. The unique cluster IDs of the most relevant clusters according to our scoring algorithm are labelled.

Table 2. Counts of terms where annotators independently agreed by category.

Category	Count
Affect/Mood	6
Appearance/Behaviour	78
Cognition	6
Insight	2
Mood/Anxiety/Affect	26
Other	102
Perception	9
Personality	23
Speech	63
Thought	22

document corpus wherein the term was detected; the category assigned to the term by each of our clinical annotators, and the SNOMED CT ID code for each term, where mapping was possible.

The most frequently detected concept mentions include 'affect' (detected in 91% of patients), 'eye contact' (85%), 'hallucinations' (85%), 'delusions' (83%) and 'rapport' (81%). Other concepts follow a long tailed distribution, with mentions of the top 407 concepts found in at least 100 unique patient records.

Regarding SNOMED CT mapping, it was possible to suggest direct mappings for 177 concepts and to suggest synonymy or partial mapping for another 53 concepts. This left a remaining 327 concepts that did not appear to be referenced in SNOMED CT, of which 106 were classified as belonging to a substantive symptom category by independent curation.

Figure 4 visualises the top 20% most frequent terms by appearance in unique patient records, where annotators agreed and were not classified as our 'other' grouping.

Owing to the difficulty of the IAA and categorisation task, an extended analysis of the top 40% most frequent terms by appearance in unique patient records, irrespective of IAA and categorisation is provided in Supplementary Figure 1.

In this project, we sought to explore SMI symptomatology and other language constructs as expressed by clinicians in their

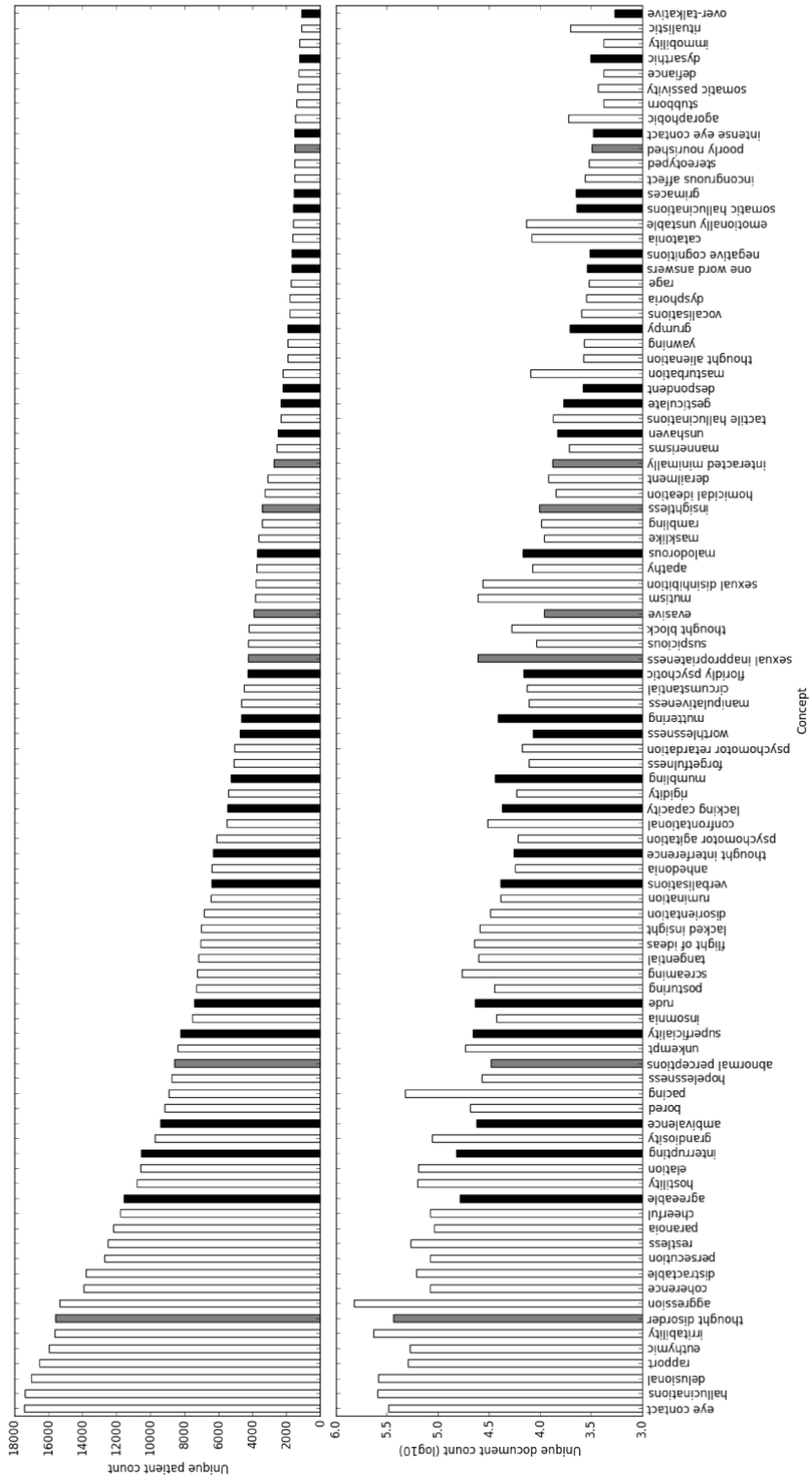


Figure 4. Frequency of terms across all SMI documents in CRIS. White bars represent concepts that were found to exist in SNOMED CT. Grey bars represent partial/uncertain matches, or novel synonyms of existing SNOMED CT concepts. Black bars represent concepts with no SNOMED mapping.

own words, using more than ten years of observations made during real-world clinician/patient interactions from more than 20 000 unique SMI cases. Within the context of a large mental healthcare provider, the results of our vocabulary curation efforts suggest that psychiatrists make use of a wide range of vocabulary to describe detailed symptomatic observations.

Many of the curated entities where both annotators agreed upon a substantive category map directly to preferred terms or synonyms of well known symptomatology constructs as described in SNOMED CT. Reassuringly, many of most frequently encountered entities as represented by unique patient count are represented in SNOMED CT, suggesting that SNOMED CT offers a reasonable coverage of what clinicians deem to be the most salient features of a psychiatric examination.

Nevertheless, our work produces evidence to suggest that many suitable synonyms are currently missing from SNOMED CT symptom entities. For instance, 'aggression' is commonly observed in SMI patients. Our results indicate that this construct might also be referred to by adjectives and phrases such as 'combatative' [*sic*], 'assaultative' [*sic*], 'truculent', 'stared intimidatingly' and 'stared menacingly', amongst others. Similarly, direct synonyms of 'paranoia' might include 'suspiciousness', 'mistrustful' and 'conspirational'[*sic*].

In addition, many of the curated constructs appear to reflect more granular observations of known symptomatology. For example, the PANSS utilises a 30-point scale of different symptomatology constructs. Specifically regarding abnormal speech, the PANSS provide guidance amounting to the high level clinical scrutiny of 'lack of spontaneity & flow of conversation'. However, clinical expressions of speech within our dataset suggest around 68 distinct states, including 'making animal noises', 'staccato quality', 'easily interruptible', 'prosody' and 'silently mouthing'.

We note the occurrence of several constructs that defy classification under existing schemas of SMI symptomatology, such as behaviours of 'over politeness', 'over complimentary', 'spending recklessly' and 'shadow boxing'. The clinical interpretation of such entities is a non-trivial exercise, and is out of scope for this piece. Nevertheless, word embedding models may offer the potential to gain insight into potentially novel symptomatology constructs observed from real-world clinician/patient interactions. Future work might explore the context for such constructs in more detail.

The emergence of such diverse language in turn has implications for how SNOMED CT might be implemented within an SMI context, raising the question of whether such gaps represent significant barriers to the use of SNOMED CT as a phenotyping resource. The issue of SNOMED CT's sufficiency in this context has previously been raised for other areas, such as rare disease[36], psychological assessment instruments[37] and histopathology findings[38]. However, in fairness, SNOMED CT is not a static resource, but an international effort dependent on the contributions of researchers. Perhaps a more pertinent question for the future development of SNOMED CT concerns balancing its objective to be a comprehensive terminology of clinical language (capable of facilitating interoperability and modelling deep phenotypes within disparate healthcare organisations across the globe) and the overwhelming complexity it would need to encompass in order to not constrain its users. Certainly, at more than 300 000 entities in its current incarnation, its size already presents problems in biomedical applications[39].

Limitations and future work

On the basis that manifestations of symptoms are the result of abnormal mental processes, novel symptom entities possibly represent observations of clinical significance. However, one particular complication in validating the clinical utility of novel symptomatology constructs with historic routinely recorded notes arises from systemic biases in EHR data. Specifically, the breadth and depth of symptomatic reporting is likely to be highly variable for a number of reasons. For instance, established symptoms as defined by current diagnostic frameworks are likely to be preferentially recorded, as clinicians are mandated to capture such entities in their assessments. On the other hand, constructs that fall outside of such frameworks may only be recorded as tangential observations made during patient/clinician interactions. Regardless of whether they are observed or not, without an established precedent of their clinical utility, they may be subject to random variation as to whether they are documented in a patient's notes. This is borne out by the tendency of SNOMED CT-ratified concepts to appear more frequently in unique documents compared to our derived expressions. The validation of new symptoms from historic data is therefore something of a 'chicken and egg' situation, a widely-discussed limitation of the reuse of EHR data[40,41]. Nevertheless, our frequency analysis of our discovered constructs suggests that there is evidence that many are observed often enough to warrant their consideration within an expanded framework. Similarly, older frameworks with a limited scope of symptomatic expression were likely designed with pragmatic constraints around speed and reproducibility of assessment in mind. However, modern technology allows for a far greater scope of data capture and validation going forward, creating opportunities to develop new frameworks that maximise the value of psychiatric assessment. Future work in this domain might seek statistical validation via randomised experimental design, as opposed to observational study.

Our work suggests an approximate correlation between patient and document count, such that intra and inter patient symptomatological clinical language usage varies relatively consistently. However, some notable exceptions to this correlation (i.e. with a higher document level frequency to patient record level frequency) include 'aggression', 'pacing', 'sexual inappropriateness', 'sexual disinhibition' and 'mutism'. Further work might seek to study these effects in greater detail, to uncover whether they represent a systemic bias in how such concepts are represented in the EHR.

The results of our IAA exercise between two experienced psychiatrists suggested a moderate level of agreement in categorising the newly identified constructs. Given that this annotation exercise did not provide any context beyond the term, and that the nature of SMI symptom observation is somewhat subjective, perhaps it is to be expected that agreement was not higher.

As suggested during peer review, providing a concordance of some of the instances of each term, along with expert panel discussion and engagement with international collaborative efforts in SMI research may prove valuable in seeking more formal definitions of the identified concepts.

Our method for vocabulary building produced nearly 1 million terms. A manual annotation of this list may have resulted in further discoveries, although would have been intractable in practical terms. To reduce the volume of terms taken forward for curation, we employed a word embedding model with a clustering algorithm. With our cluster scoring methodology that makes use of existing domain knowledge, we were able to successfully produce meaningful clusters of terms reflecting the semantics of SMI symptomatology. However, as with many unsupervised tasks, it is difficult to determine whether an optimal solution has been achieved. In particular, the emergence of three 'symptom' clusters instead of one indicates sub-optimal localisation of symptom constructs in vector space. Addressing such a problem is multifaceted. For technical reasons, only a single epoch of training was possible in this exercise. Additional epochs would likely contribute to better cluster definition, in turn allowing us to reduce the value of our k parameter. In addition, spell checking and collapsing terms into their root forms may also have assisted. However, the latter may have also created new word sense disambiguation problems if common, symptom-like morphemes also appear in nonsymptomatological assessment contexts.

After clustering, a two stage manual curation of more than 20 000 terms was necessary. Methods that produce a smaller vocabulary might conceivably reduce annotator burden. This might include the use of spell checkers and stemming/lemmatisation to correct and normalise tokens, at the risk of introducing new issues associated with morphological forms in word embedding model building. For this attempt, we took the conscious decision to make as few assumptions about the underlying structure of the data as possible.

During peer review, it was suggested that recent advancements in topic modelling approaches may be relevant to our work. Many groups have sought to combine the popular technique of Latent Dirichlet Allocation (LDA)[42] with word embedding models to derive appropriate terminology for a given topic[43–45]. For instance, Nguyen *et al.*[46] propose an extension of LDA that makes use of a word embedding model trained on a very large corpus of text to improve the performance of topic coherence modelling on several datasets. Future work might seek to explore such techniques, and (assuming regulatory barriers can be overcome), the potential of creating word embedding models from very large clinical text corpora by combining data with other care organisations.

Conclusions

Evidence-based mental health has long sought to produce disease model definitions that are both valid, in the sense they represent useful clinical representations that can inform treatment, and reliable, in that they can be consistently applied by different clinicians to achieve the same outcomes. In practice this has proven difficult, due to the often subjective nature of psychiatric examination/phenotyping and insufficient knowledge about the underlying mechanisms of disorders such as SMI. Here, we demonstrate that clinical staff make use of a diverse vocabulary in the course of their interactions with patients. This vocabulary often references findings that are not represented in SNOMED CT, raising questions about whether clinicians should observe the constraints of SNOMED CT or whether SNOMED CT should incorporate greater flexibility to reflect the nature of mental health. It is outside the scope of this work to explore how the granularity of symptom-based phenotyping affects patient outcomes, although the possibility of offering a fully realised picture of symptom manifestation may prove valuable in future endeavours of precision medicine.

Data availability

The CRIS dataset is a pseydonymised and de-identified case registrar of electronic health records of the SLaM NHS Trust. It operates under a security model that does not allow for open publication of raw data. However, access can be granted for research use cases under a patient-led security model. For further information and details on the application process, please contact cris.administrator@kcl.ac.uk or visit the website: https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/. Alternatively, you may write to the CRIS team at:

PO Box 92 Institute of Psychiatry, Psychology & Neuroscience at King's College London 16 De Crespigny Park London SE5 8AF

Example code used in this analysis is available at: https://github.com/RichJackson/clustering_w2v

Competing interests

RJ and RS have received research funding from Roche, Pfizer, J&J and Lundbeck.

Grant information

This paper represents independent research funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. RP has received support from a Medical Research Council (MRC) Health Data Research UK Fellowship and a Starter Grant for Clinical Lecturers (SGL015/1020) supported by the Academy of Medical Sciences, The Wellcome Trust, MRC, British Heart Foundation, Arthritis Research UK, the Royal College of Physicians and Diabetes UK. SV is supported by the Swedish Research Council (2015-00359), Marie Sklodowska Curie Actions, Cofund, Project INCA 600398.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary material

Supplementary File 1: This file contains all of the 557 terms taken forward for expert annotation. It includes SNOMED mappings where possible, unique document and patient counts within the corpus, and the annotations provided by RP and RS.

Click here to access the data.

Supplementary Figure 1: This file is an expanded visualisation of the frequency analysis figure contained in the main manuscript, with the agreement and nonsubstantive 'other' classification restrictions lifted.

Click here to access the data.

References

1. Amberger J, Bocchini C, Hamosh A: **A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®).** *Hum Mutat.* 2011; **32**(5): 564–567. ISSN 10597794.
PubMed Abstract | Publisher Full Text
2. Mirnezami R, Nicholson J, Darzi A: **Preparing for precision medicine.** *N Engl J Med.* 2012; **366**(6): 489–491. ISSN 0028-4793, 1533-4406.
PubMed Abstract | Publisher Full Text
3. Robinson PN: **Deep phenotyping for precision medicine.** *Hum Mutat.* 2012; **33**(5): 777–780. ISSN 10597794.
PubMed Abstract | Publisher Full Text
4. Pathak J, Kho AN, Denny JC: **Electronic health records-driven phenotyping: challenges, recent advances, and perspectives.** *J Am Med Inform Assoc.* 2013; **20**(e2): e206–11.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Castro VM, Minnier J, Murphy SN, *et al.*: **Validation of electronic health record phenotyping of bipolar disorder cases and controls.** *Am J Psychiatry.* 2015; **172**(4): 363–372. ISSN 0002-953X, 1535-7228.
PubMed Abstract | Publisher Full Text | Free Full Text
6. NATIONAL INFORMATION BOARD: **Personalised Health and Care 2020.** 2014.
Reference Source
7. Lee D, Cornet R, Lau F, *et al.*: **A survey of SNOMED CT implementations.** *J Biomed Inform.* 2013; **46**(1): 87–96. ISSN 15320464.
PubMed Abstract | Publisher Full Text
8. Barnes M: **Lessons learned from the implementation of clinical messaging systems.** *AMIA Annu Symp Proc.* 2007; 36–40. ISSN 1942-597X.
PubMed Abstract | Free Full Text
9. **The future of healthcare informatics: it is not what you think.** *Glob Adv Health Med.* 2012; **1**(4): 5–6. ISSN 2164-957X, 2164-9561.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Gordon D: **Merging multiple institutions: Information architecture problems and solutions.** *Proc AMIA Symp.* 1999; 785–789. ISSN 1531-605X.
PubMed Abstract | Free Full Text
11. Freedman R, Lewis DA, Michels R, *et al.*: **The initial field trials of DSM-5: new blooms and old thorns.** *Am J Psychiatry.* 2013; **170**(1): 1–5. ISSN 0002-953X, 1535-7228.
PubMed Abstract | Publisher Full Text
12. Kendell R, Jablensky A: **Distinguishing between the validity and utility of psychiatric diagnoses.** *Am J Psychiatry.* 2003; **160**(1): 4–12. ISSN 0002-953X, 1535-7228.
PubMed Abstract | Publisher Full Text
13. Chmielewski M, Bagby RM, Markon K, *et al.*: **Openness to experience, intellect, schizotypal personality disorder, and psychoticism: resolving the controversy.** *J Pers Disord.* 2014; **28**(4): 483–99. ISSN 1943-2763.
PubMed Abstract | Publisher Full Text
14. Adam D: **Mental health: On the spectrum.** *Nature.* 2013; **496**(7446): 416–418. ISSN 0028-0836, 1476-4687.
PubMed Abstract | Publisher Full Text
15. Cross-Disorder Group of the Psychiatric Genomics Consortium: **Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis.** *Lancet.* 2013; **381**(9875): 1371–1379. ISSN 01406736.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Kay SR, Fiszbein A, Opler LA: **The positive and negative syndrome scale (PANSS) for schizophrenia.** *Schizophr Bull.* 1987; **13**(2): 261–76. ISSN 0586-7614. Cited by 8221.
PubMed Abstract | Publisher Full Text
17. Kirkpatrick B, Strauss GP, Nguyen L, *et al.*: **The brief negative symptom scale: psychometric properties.** *Schizophr Bull.* 2010; **37**(2): 300–305. IISSN 0586-7614, 1745-1701. Cited by 0000.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Liu H, Aronson AR, Friedman C: **A study of abbreviations in MEDLINE abstracts.** *Proc AMIA Symp.* 2002; 464–468. ISSN 1531-605X.
PubMed Abstract | Free Full Text
19. Henriksson A, Conway M, Duneld M, *et al.*: **Identifying synonymy between SNOMED clinical terms of varying length using distributional analysis of electronic health records.** *AMIA Annu Symp Proc.* 2013; **2013**: 600–609. ISSN 1942-597X.
PubMed Abstract | Free Full Text
20. Krauthammer M, Nenadic G: **Term identification in the biomedical literature.** *J Biomed Inform.* 2004; **37**(6): 512–526. ISSN 15320464.
PubMed Abstract | Publisher Full Text
21. Boksa P: **A way forward for research on biomarkers for psychiatric disorders.** *J Psychiatry Neurosci.* 2013; **38**(2): 75–55. ISSN 11804882.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Jackson RG, Patel R, Jayatilleke N, *et al.*: **Natural language processing to extract symptoms of severe mental illness from clinical text: The Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project.** *BMJ Open.* 2017; **7**(1): e012012. ISSN 2044-6055, 2044-6055.
PubMed Abstract | Publisher Full Text | Free Full Text
23. McGorry PD: **The next stage for diagnosis: Validity through utility.** *World Psychiatry.* 2013; **12**(3): 213–215. ISSN 17238617.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Perera G, Broadbent M, Callard F, *et al.*: **Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: Current status and recent enhancement of an Electronic Mental Health Record-derived data resource.** *BMJ Open.* 2016; **6**(3): e008721. ISSN 2044-6055, 2044-6055.
PubMed Abstract | Publisher Full Text | Free Full Text
25. Bird S, Klein E, Loper E: **Natural Language Processing with Python.** O'Reilly, Beijing; Cambridge [Mass.], 1st ed edition, 2009. ISBN 978-0-596-51649-9. OCLC: ocn301885973.
Reference Source
26. Řehůřek R, Sojka P: **Software Framework for Topic Modelling with Large Corpora**. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.* 2010; 45–50. Valletta, Malta, ELRA.
Publisher Full Text
27. Mikolov T, Sutskever I, Chen K, *et al.*: **Distributed representations of words and phrases and their compositionality.** *Adv Neural Inf Process Syst.* 2013; 3111–3119.
Reference Source
28. Harris ZS: **Distributional Structure.** *WORD.* 1954; **10**(2–3): 146–162. ISSN 0043-7956, 2373-5112.
Publisher Full Text
29. Mikolov T, Chen K, Corrado G, *et al.*: **Efficient estimation of word representations in vector space.** *arXiv preprint arXiv: 1301.3781.* 2013.
Reference Source
30. Pakhomov SV, Finley G, McEwan R, *et al.*: **Corpus domain effects on distributional semantic modeling of medical terms.** *Bioinformatics.* 2016; **32**(23): 3635–3644. ISSN 1367-4803, 1460-2059.
PubMed Abstract | Publisher Full Text | Free Full Text
31. Rong X: **Word2vec parameter learning explained.** *arXiv preprint arXiv: 1411.2738.* 2014.
Reference Source
32. Pedregosa F, Varoquaux G, Gramfort A, *et al.*: **Scikit-learn: Machine Learning in**

Python. *J Mach Learn Res.* 2011; **12**: 2825–2830.
Reference Source

33. Kodinariya TM, Makwana PR: **Review on determining number of Cluster in K-Means Clustering.** *Int J.* 2013; **1**(6): 90–95.
Reference Source

34. Harrison PJ, Cowen P, Burns T, *et al.*: **Shorter Oxford book of psych.** In *Shorter Oxford Textbook of Psychiatry.* Oxford University Press, Oxford, seventh edition edition, 2018; 44. ISBN 978-0-19-874743-7.

35. Cohen J: **A Coefficient of Agreement for Nominal Scales.** *Educ Psychol Meas.* 1960; **20**(1): 37–46. ISSN 0013-1644, 1552-3888.
Publisher Full Text

36. Sollie A, Sijmons RH, Lindhout D, *et al.*: **A new coding system for metabolic disorders demonstrates gaps in the international disease classifications ICD-10 and SNOMED-CT, which can be barriers to genotype-phenotype data sharing.** *Hum Mutat.* 2013; **34**(7): 967–973. ISSN 10597794.
PubMed Abstract | Publisher Full Text

37. Ranallo PA, Adam TJ, Nelson KJ, *et al.*: **Psychological assessment instruments: a coverage analysis using SNOMED CT, LOINC and QS terminology.** *AMIA Annu Symp Proc.* 2013; **2013**: 1333–1340. ISSN 1942-597X.
PubMed Abstract | Free Full Text

38. Campbell WS, Campbell JR, West WW, *et al.*: **Semantic analysis of SNOMED CT for a post-coordinated database of histopathology findings.** *J Am Med Inform Assoc.* 2014; **21**(5): 885–892. ISSN 1067-5027, 1527-974X.
PubMed Abstract | Publisher Full Text | Free Full Text

39. López-García P, Schulz S: **Can SNOMED CT be squeezed without losing its shape?** *J Biomed Semantics.* 2016; **7**(1): 56. ISSN 2041-1480.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Weiskopf NG, Weng C: **Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research.** *J Am Med Inform Assoc.* 2013; **20**(1): 144–151. ISSN 1067-5027, 1527-974X.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Chan KS, Fowles JB, Weiner JP: **Review: electronic health records and the reliability and validity of quality measures: a review of the literature.** *Med Care Res Rev.* 2010; **67**(5): 503–527. ISSN 1077-5587, 1552-6801.
PubMed Abstract | Publisher Full Text

42. Blei DM, Ng AY, Jordan MI: **Latent dirichlet allocation.** *J Mach Learn Res.* 2003; **3**: 993–1022.
Reference Source

43. Cao Z, Li S, Liu Y, *et al.*: **A Novel Neural Topic Model and Its Supervised Extension.** In *AAAI.* 2015; 2210–2216.
Reference Source

44. Hinton GE, Salakhutdinov RR: **Replicated softmax: An undirected topic model.** In *Adv Neural Inf Process Syst.* 2009; 1607–1614.
Reference Source

45. Srivastava N, Salakhutdinov RR, Hinton GE: **Modeling documents with deep boltzmann machines**. *arXiv preprint arXiv:1309.6865*, 2013.
Reference Source

46. Nguyen DQ, Billingsley R, Du L, *et al.*: **Improving topic models with latent feature word representations.** *Trans Assoc Comput Linguist.* 2015; **3**: 399–313.
Reference Source

5.1.1 Supplementary File 1

Concept	SNOMED Code	SNOMED_comment	Documents	Patients	RP classifications	RS classifications
abnormal perceptions	225453006	match with new synonym	30238	8584	Perception	Perception
acrimonious		new entity	736	400	Appearance/Behaviour	Personality
adaptable	286804000	direct match (finding)	20234	6588	Other	Other
affect	416383008	direct match (finding)	441454	18710	Other	Other
aggression	61372001	direct match (finding)	659153	15351	Appearance/Behaviour	Appearance/Behaviour
agitation	24199005	direct match (finding)	401979	16115	Appearance/Behaviour	Other
agoraphobic	70691001	direct match (disorder)	5224	1471	Mood/Anxiety/Affect	Mood/Anxiety/Affect
agreeable		new entity	61289	11577	Appearance/Behaviour	Appearance/Behaviour
alertness	365933000	direct match (finding)	1645	1060	Appearance/Behaviour	Appearance/Behaviour
alexithymic	34413007	direct match (finding)	184	94	Personality	Other
alienation	247695002	direct match (finding)	5703	2989	Other	Thought
aloof	285847005	direct match (finding)	1854	1065	Personality	Other
ambivalence		new entity	41700	9411	Appearance/Behaviour	Appearance/Behaviour
amicable		new entity	6367	2982	Appearance/Behaviour	Personality
amotivational	26413003	match with new synonym	2708	1435	Appearance/Behaviour	Mood/Anxiety/Affect
anankastic traits		new entity	39	23	Personality	Personality
anhedonia	28669007	direct match (finding)	17475	6403	Affect/Mood	Affect/Mood
anomalous experiences		new entity	434	249	Other	Other
answer questions monosyllabically		new entity	77	60	Speech	Speech
answering irrelevantly	9288000	match with new synonym	73	51	Speech	Speech
anti-authoritarian attitudes		new entity	118	39	Personality	Personality
apathy	20602000	direct match (finding)	11766	3749	Affect/Mood	Affect/Mood
aphasia	87486003	direct match (finding)	357	169	Speech	Speech

apnoea	1023001	direct match (finding)	4836	686	Other	Other
argumentative	9182005	direct match (finding)	44117	6887	Appearance/Behaviour	Mood/Anxiety/Affect
articulation deficits	286296002	match with new synonym	27	18	Speech	Speech
asocial	76991007	direct match (finding)	332	217	Appearance/Behaviour	Personality
aspontaneous		new entity	77	49	Other	Other
assaultative		new entity	7253	1990	Appearance/Behaviour	Other
ataxic	13628000	partial match	641	356	Other	Other
auditory command hallucinations	45150006	direct match (finding)	1596	738	Perception	Perception
autobiographical memory	283996004	direct match (finding)	97	69	Cognition	Cognition
avoidance behaviours	284489004	direct match (finding)	578	420	Mood/Anxiety/Affect	Other
avoided answering		new entity	954	713	Appearance/Behaviour	Speech
babbling	258116008	direct match (finding)	797	367	Speech	Speech
babyish voice		new entity	32	16	Speech	Speech
barricading		new entity	7509	1795	Other	Other
belligerence		new entity	1426	836	Appearance/Behaviour	Appearance/Behaviour
beseeching		new entity	36	29	Other	Other
blank expression		new entity	621	427	Appearance/Behaviour	Mood/Anxiety/Affect
bleak outlook		new entity	62	50	Mood/Anxiety/Affect	Mood/Anxiety/Affect
bombastic		new entity	52	30	Speech	Appearance/Behaviour
bored	83765003	direct match (finding)	48067	9176	Appearance/Behaviour	Appearance/Behaviour
boundried		new entity	78	52	Other	Other
breathless	267036007	direct match (finding)	14811	3994	Other	Other
brusque		new entity	440	331	Appearance/Behaviour	Appearance/Behaviour
buzzy		new entity	274	166	Mood/Anxiety/Affect	Other

callousness	286734003	direct match (finding)	1467	599	Personality	Personality
cantankerous		new entity	170	120	Appearance/Behaviour	Personality
capricious		new entity	121	89	Appearance/Behaviour	Personality
catalepsy		new entity	53	29	Appearance/Behaviour	Appearance/Behaviour
catastrophic thinking	285247003	match with new synonym	319	225	Thought	Thought
catatonia	247917007	direct match (finding)	11946	1652	Appearance/Behaviour	Appearance/Behaviour
cheerful	112080002	direct match (finding)	118509	11744	Mood/Anxiety/Affect	Mood/Anxiety/Affect
circuitous		new entity	137	105	Thought	Other
circumferential	255593009	Direct match (qualifier)	472	292	Thought	Other
circumscribed	263706005	Direct match (qualifier)	710	488	Other	Other
circumstantial	18343006	direct match (finding)	13490	4479	Speech	Speech
clang associations	88293003	direct match (finding)	1101	594	Thought	Speech
clapping hands		new entity	175	109	Appearance/Behaviour	Speech
claustrophobic	19887002	direct match (finding)	50	41	Mood/Anxiety/Affect	Other
cleanliness	248155000	Direct match (observable entity)	6953	3106	Other	Other
clipped answers	289195008	match with new synonym	43	27	Speech	Speech
cluster headaches	193031009	direct match (disorder)	83	30	Other	Other
coerced		new entity	3841	2038	Other	Other
cognisant		new entity	200	164	Other	Other
cognitive biases		new entity	256	165	Other	Other
cognitive dissonance		new entity	48	39	Other	Other
coherence	284596004	direct match (finding)	120008	13900	Speech	Speech
collusive		new entity	42	35	Other	Other
combatative		new entity	51	43	Appearance/Behaviour	Mood/Anxiety/Affect

commanding hallucination	78595002	direct match (finding)	28577	5612	Hallucination	Perception
concrete thinking	71573006	direct match (finding)	1852	962	Thought	Thought
confabulating	17842005	direct match (finding)	1528	647	Cognition	Speech
conflictual interpersonal style		new entity	25	18	Personality	Personality
confrontational	284662003	direct match (finding)	32582	5500	Appearance/Behaviour	Appearance/Behaviour
conjuring		new entity	118	94	Other	Other
consequential thinking		new entity	219	116	Other	Thought
conspiratorially		new entity	886	527	Other	Other
constant interruptions		new entity	427	343	Speech	Speech
constrictive		new entity	33	28	Other	Other
convoluted	47573009	Direct match (qualifier)	948	654	Thought	Other
coquettish		new entity	27	21	Appearance/Behaviour	Appearance/Behaviour
counter transference	224982006	direct match (finding)	293	70	Other	Other
crying hysterically	271951008	match with new synonym	179	113	Appearance/Behaviour	Appearance/Behaviour
culturally normative		new entity	424	292	Other	Other
deceptive		new entity	258	186	Other	Other
decreased blink	416721005	direct match (finding)	29	19	Appearance/Behaviour	Appearance/Behaviour
defeatist		new entity	83	69	Personality	Personality
defiance	248039009	direct match (finding)	2368	1267	Appearance/Behaviour	Appearance/Behaviour
deflecting		new entity	1194	804	Other	Other
delusional	2073000	direct match (finding)	381772	17002	Thought	Thought
demanding	284503003	direct match (finding)	194890	12233	Appearance/Behaviour	Other
demure		new entity	164	134	Appearance/Behaviour	Other
denigratory		new entity	87	68	Other	Other

derailment	65135009	direct match (finding)	8221	3124	Speech	Speech
derealization	40806005	direct match (finding)	1972	1044	Mood/Anxiety/Affect	Mood/Anxiety/Affect
derisive		new entity	62	51	Appearance/Behaviour	Other
derogative		new entity	30098	6437	Appearance/Behaviour	Other
despondent		new entity	3806	2214	Mood/Anxiety/Affect	Mood/Anxiety/Affect
destroying furniture	284661005	match with new synonym	248	111	Appearance/Behaviour	Other
desultory		new entity	40	31	Appearance/Behaviour	Other
deviant sexual arousal		new entity	18	14	Appearance/Behaviour	Other
dichotomous thinking		new entity	126	107	Personality	Personality
didactic		new entity	80	63	Other	Other
dietary neglect		new entity	41	22	Appearance/Behaviour	Other
different accents		new entity	319	176	Speech	Speech
directionless		new entity	81	64	Other	Other
disagreeable		new entity	269	215	Appearance/Behaviour	Other
disarrayed		new entity	2892	1400	Other	Other
disciplinarian		new entity	412	166	Personality	Personality
disconfirmatory evidence		new entity	48	31	Other	Other
disconsolate		new entity	58	47	Mood/Anxiety/Affect	Other
discourteous		new entity	107	75	Appearance/Behaviour	Other
discursive		new entity	444	287	Other	Speech
disempowered	225793008	direct match (finding)	565	426	Other	Other
disgruntled		new entity	1677	1149	Appearance/Behaviour	Other
disillusioned		new entity	691	476	Thought	Thought
dismissive towards		new entity	733	493	Other	Other

disorientation	62476001	direct match (finding)	30593	6834	Cognition	Cognition
dissociated		new entity	9804	2244	Other	Other
distorted thoughts		new entity	158	119	Thought	Thought
distractable	163616009	direct match (finding)	162979	13807	Appearance/Behaviour	Appearance/Behaviour
domineering	286787003	direct match (finding)	29720	6670	Appearance/Behaviour	Personality
downhearted		new entity	69	60	Mood/Anxiety/Affect	Mood/Anxiety/Affect
dresses completely inappropriate	248162009	match with new synonym	2141	983	Appearance/Behaviour	Other
drilling noise		new entity	92	67	Other	Other
drooling	62718007	direct match (finding)	3036	1336	Other	Appearance/Behaviour
drossy		new entity	75	67	Sleep	Other
dunning		new entity	1594	556	Other	Other
durable		new entity	103	64	Other	Other
dysarthic		new entity	3210	1253	Speech	Speech
dysdiadokinesis		new entity	133	104	Other	Other
dysfluent		new entity	17	11	Speech	Speech
dysmetria	32566006	direct match (finding)	173	113	Cognition	Cognition
dysphasia	20301004	direct match (finding)	1371	640	Speech	Speech
dysphonic		new entity	183	103	Speech	Speech
dysphoria	30819006	direct match (finding)	3523	1801	Mood/Anxiety/Affect	Mood/Anxiety/Affect
easily interruptible		new entity	207	170	Speech	Speech
easily redirected		new entity	307	224	Other	Appearance/Behaviour
echolalia	64712007	direct match (finding)	1588	620	Speech	Speech
echopraxia	33184005	direct match (finding)	274	137	Appearance/Behaviour	Appearance/Behaviour
ecstatic	286582007	direct match (finding)	345	265	Mood/Anxiety/Affect	Mood/Anxiety/Affect

edginess		new entity	277	225	Appearance/Behaviour	Appearance/Behaviour
egodystonic	52813007	direct match (finding)	302	192	Thought	Other
egotistical		new entity	55	39	Personality	Personality
elaborative		new entity	202	142	Other	Speech
elation	34822003	direct match (finding)	155348	10585	Affect/Mood	Affect/Mood
elusive		new entity	1251	831	Other	Other
emotionally drained	225650001	match with new synonym	368	280	Other	Other
emotionally immature	90415008	direct match (finding)	150	88	Personality	Personality
emotionally numb		new entity	288	197	Mood/Anxiety/Affect	Mood/Anxiety/Affect
emotionally unstable	191765005	direct match (disorder)	13557	1613	Personality	Personality
erotomanic	280949006	direct match (finding)	950	238	Thought	Other
euphoria	85949006	direct match (finding)	1061	668	Affect/Mood	Affect/Mood
euthymic	82248001	direct match (finding)	188290	15953	Mood/Anxiety/Affect	Mood/Anxiety/Affect
evangelising		new entity	409	144	Other	Other
evasive	274651006	match with new synonym	9065	3919	Appearance/Behaviour	Appearance/Behaviour
exasperating		new entity	1657	1224	Other	Other
excess drowsiness	271782001	direct match (finding)	936	685	Sleep	Appearance/Behaviour
experienced vivid nightmares	419145002	match with new synonym	32	23	Sleep	Other
exploitative		new entity	951	565	Other	Other
expressed preoccupying thoughts		new entity	226	174	Thought	Thought
expressionless	248149005	direct match (finding)	1172	745	Appearance/Behaviour	Appearance/Behaviour
extroverted	224954004	direct match (finding)	558	309	Personality	Personality
exuberantly		new entity	43	37	Appearance/Behaviour	Other
eye contact	412786000	direct match (finding)	305869	17409	Appearance/Behaviour	Appearance/Behaviour

facetious		new entity	303	202	Appearance/Behaviour	Personality
facial asymmetry		new entity	681	469	Appearance/Behaviour	Appearance/Behaviour
facially bright		new entity	204	125	Appearance/Behaviour	Appearance/Behaviour
factuous		new entity	30	27	Other	Other
fanciful		new entity	74	64	Other	Other
fastidious		new entity	233	143	Appearance/Behaviour	Personality
fatalistic		new entity	233	179	Mood/Anxiety/Affect	Personality
fatigued	84229001	direct match (finding)	3040	1760	Sleep	Mood/Anxiety/Affect
fatuous affect	247654000	match with new synonym	252	141	Mood/Anxiety/Affect	Mood/Anxiety/Affect
fervent		new entity	267	217	Other	Other
firmly rooted		new entity	13	11	Other	Other
fixated		new entity	23628	5559	Appearance/Behaviour	Other
flambouyantly		new entity	432	225	Appearance/Behaviour	Other
flash backs	30871003	direct match (finding)	1110	631	Perception	Thought
fleetingly	247765001	partial match	559	423	Other	Other
flight of ideas	28810003	direct match (finding)	43848	7062	Speech	Speech
flippantly		new entity	91	80	Appearance/Behaviour	Other
flirtatious		new entity	2264	1043	Appearance/Behaviour	Appearance/Behaviour
floridly psychotic		new entity	14600	4305	Thought	Thought
forgetfulness	55533009	direct match (finding)	12720	5092	Cognition	Cognition
formulaic		new entity	31	27	Thought	Other
fragile ego		new entity	836	580	Personality	Personality
free-floating anxiety	81350009	direct match (finding)	449	290	Mood/Anxiety/Affect	Mood/Anxiety/Affect
frequent anger	274951009	partial match	41	31	Appearance/Behaviour	Mood/Anxiety/Affect

fretful		new entity	214	163	Mood/Anxiety/Affect	Personality
friable		new entity	54	39	Other	Other
frigidity	62607004	direct match (finding)	38	34	Other	Personality
frivolous		new entity	175	135	Appearance/Behaviour	Other
frosty		new entity	352	285	Appearance/Behaviour	Other
futile	225473001	direct match (finding)	2231	1452	Other	Other
garrulous		new entity	788	466	Speech	Speech
gauche		new entity	56	43	Appearance/Behaviour	Other
gaze avoidant	412786000	match with new synonym	203	128	Appearance/Behaviour	Appearance/Behaviour
gesticulate		new entity	5838	2327	Appearance/Behaviour	Appearance/Behaviour
ghost-like		new entity	3746	1427	Appearance/Behaviour	Other
gibberish	91646005	partial match	687	417	Speech	Speech
giddy	271789005	direct match (finding)	995	556	Other	Other
giggling inappropriately	247985007	match with new synonym	2403	1000	Mood/Anxiety/Affect	Mood/Anxiety/Affect
glib		new entity	423	185	Appearance/Behaviour	Other
gloomy		new entity	1199	768	Mood/Anxiety/Affect	Mood/Anxiety/Affect
grandiosity	247783009	direct match (finding)	113036	9739	Thought	Thought
great ambivalence		new entity	18	16	Appearance/Behaviour	Other
grimaces		new entity	4463	1551	Appearance/Behaviour	Appearance/Behaviour
grumpy		new entity	5106	1925	Mood/Anxiety/Affect	Mood/Anxiety/Affect
grunting noise		new entity	2771	881	Speech	Speech
gun gestures		new entity	550	148	Appearance/Behaviour	Appearance/Behaviour
gustatory hallucination	29139005	direct match (finding)	634	356	Perception	Perception
guttural noises		new entity	59	22	Speech	Speech

haggard	248186006	direct match (finding)	154	118	Appearance/Behaviour	Appearance/Behaviour
half hearted		new entity	435	352	Other	Other
hallucinations	7011001	direct match (finding)	387103	17368	Perception	Perception
hand-wringing	41996009	direct match (finding)	229	160	Appearance/Behaviour	Appearance/Behaviour
harmonious		new entity	370	258	Other	Other
helplessness	33300005	direct match (finding)	9379	4283	Other	Mood/Anxiety/Affect
hesistant		new entity	615	518	Speech	Other
highly strung		new entity	180	126	Personality	Personality
hissing sounds		new entity	150	81	Speech	Other
histrionically	248023002	direct match (finding)	1798	742	Personality	Other
hoarding	248025009	direct match (finding)	10879	2626	Mood/Anxiety/Affect	Other
homesick		new entity	562	340	Other	Other
homicidal ideation	225450009	direct match (finding)	6926	3280	Thought	Thought
hopelessness	307077003	direct match (finding)	36816	8741	Mood/Anxiety/Affect	Mood/Anxiety/Affect
hostility	79351003	direct match (finding)	158210	10795	Appearance/Behaviour	Appearance/Behaviour
howling noises		new entity	147	50	Speech	Speech
humiliation		new entity	3005	1534	Other	Other
humming		new entity	1878	793	Speech	Speech
Hyper-religious		new entity	396	209	Thought	Thought
Hyper-sensitive	421369008	direct match (finding)	263	191	Other	Personality
Hyper-sexual	73744004	direct match (finding)	532	281	Appearance/Behaviour	Appearance/Behaviour
Hyper-vigilant	423752000	direct match (finding)	1327	771	Appearance/Behaviour	Thought
hyperthymic		new entity	264	169	Mood/Anxiety/Affect	Mood/Anxiety/Affect
hypnagogic hallucinations	44780000	direct match (finding)	134	92	Perception	Perception

hypnapompic hallucinations	69690008	direct match (finding)	209	132	Perception	Perception
hypo-mania	231496004	match with new synonym	1373	838	Mood/Anxiety/Affect	Other
hypochondrial delusions	247688006	match with new synonym	475	210	Thought	Thought
hypomanic phases	284512001	direct match (finding)	529	332	Mood/Anxiety/Affect	Other
hypothymic		new entity	212	96	Mood/Anxiety/Affect	Mood/Anxiety/Affect
immobility	67759008	direct match (finding)	2385	1238	Appearance/Behaviour	Appearance/Behaviour
implanting		new entity	135	77	Other	Other
inarticulate	286278006	match with new synonym	139	91	Speech	Speech
incongruous affect	404652005	direct match (finding)	3613	1540	Mood/Anxiety/Affect	Mood/Anxiety/Affect
increased clumsiness	7006003	direct match (finding)	166	81	Appearance/Behaviour	Appearance/Behaviour
increased talkativeness		new entity	532	338	Speech	Speech
increasingly avoidant	304869006	partial match	65	36	Mood/Anxiety/Affect	Other
increasingly confused	40917007	direct match (finding)	536	333	Cognition	Other
increasingly preoccupied	248235009	direct match (finding)	1079	621	Other	Thought
infatuated	18318000	direct match (finding)	522	267	Appearance/Behaviour	Other
inflections		new entity	371	257	Speech	Other
inoffensive		new entity	63	50	Other	Other
insightless	24340004	match with new synonym	10151	3429	Insight	Insight
insomnia	53758003	direct match (finding)	26750	7533	Affect/Mood	Affect/Mood
institutionalised	225800006	direct match (finding)	3337	1324	Other	Other
intense eye contact		new entity	2995	1508	Appearance/Behaviour	Appearance/Behaviour
interacted minimally	88598008	partial match	7543	2726	Appearance/Behaviour	Appearance/Behaviour
intermittently engaging		new entity	1006	523	Appearance/Behaviour	Appearance/Behaviour
internalisation		new entity	1480	849	Other	Other

interpersonal hypersensitivity		new entity	55	35	Personality	Personality
interrogative		new entity	1274	823	Appearance/Behaviour	Other
interrupting		new entity	66233	10547	Speech	Speech
intransigent		new entity	146	123	Personality	Personality
intrusive recollections		new entity	37	26	Thought	Thought
irascible		new entity	236	168	Mood/Anxiety/Affect	Appearance/Behaviour
irritability	55929007	direct match (finding)	424911	15609	Affect/Mood	Affect/Mood
isolation	422650009	direct match (finding)	230268	14211	Other	Other
jittery	424196004	direct match (finding)	856	566	Appearance/Behaviour	Mood/Anxiety/Affect
jocular	5240007	direct match (finding)	640	377	Appearance/Behaviour	Mood/Anxiety/Affect
jovial		new entity	21829	4999	Appearance/Behaviour	Other
knights move thinking	65135009	direct match (finding)	1267	710	Thought	Speech
lackadaisical		new entity	25	21	Appearance/Behaviour	Personality
lacked insight	24340004	direct match (finding)	38584	7018	Insight	Insight
lacking capacity		new entity	23694	5456	Cognition	Cognition
lacks assertiveness	286788008	partial match	102	69	Appearance/Behaviour	Personality
laconic		new entity	199	142	Speech	Other
language processing		new entity	22	19	Speech	Other
latency		new entity	2339	1059	Speech	Other
lethargic	214264003	direct match (finding)	19876	6740	Sleep	Other
lewd		new entity	409	208	Appearance/Behaviour	Other
licking	711610000	partial match	1144	533	Other	Other
lilting		new entity	37	29	Speech	Speech
litigious	286844007	direct match (finding)	200	113	Other	Personality

loneliness	267076002	direct match (finding)	8460	3519	Other	Thought
long latencies		new entity	445	237	Speech	Other
loquacious		new entity	169	126	Speech	Speech
low profiled		new entity	235042	8326	Appearance/Behaviour	Other
maladaptive behaviours	284499009	direct match (finding)	490	244	Personality	Personality
malapropisms		new entity	33	20	Speech	Speech
malevolent intentions		new entity	96	73	Other	Other
malodorous		new entity	14724	3711	Appearance/Behaviour	Appearance/Behaviour
manipulativeness	261935004	direct match (finding)	12805	4647	Personality	Personality
mannerisms	248026005	direct match (finding)	5170	2582	Appearance/Behaviour	Appearance/Behaviour
masklike	103606006	direct match (finding)	9120	3625	Appearance/Behaviour	Appearance/Behaviour
masturbation	17704007	direct match (finding)	12482	2193	Appearance/Behaviour	Appearance/Behaviour
meandering		new entity	192	152	Thought	Other
melodic		new entity	398	317	Speech	Other
menacing		new entity	2230	1023	Appearance/Behaviour	Mood/Anxiety/Affect
mentally fabulous		new entity	314	18	Other	Other
miaowing		new entity	22	18	Speech	Speech
mild tremulousness		new entity	67	52	Other	Appearance/Behaviour
mimicking	31271002	direct match (finding)	1970	934	Other	Appearance/Behaviour
mis-trustful		new entity	3623	1980	Appearance/Behaviour	Mood/Anxiety/Affect
misogynistic		new entity	240	114	Other	Personality
morose		new entity	394	262	Mood/Anxiety/Affect	Mood/Anxiety/Affect
mumbling		new entity	27527	5272	Speech	Speech
mutism	88052002	direct match (finding)	40389	3822	Speech	Speech

muttering		new entity	25917	4661	Speech	Speech
needs daily prompting		new entity	132	80	Other	Other
negative cognitions		new entity	3246	1686	Mood/Anxiety/Affect	Mood/Anxiety/Affect
neologisms	54501006	direct match (finding)	3604	1366	Thought	Speech
nonsensical statements		new entity	66	44	Speech	Speech
normothymic	82248001	match with new synonym	189	132	Mood/Anxiety/Affect	Mood/Anxiety/Affect
numbness	44077006	direct match (finding)	4380	2177	Other	Other
oblique references		new entity	89	52	Thought	Other
officious		new entity	52	38	Appearance/Behaviour	Personality
one word answers		new entity	3446	1694	Speech	Speech
oppressive		new entity	1550	835	Appearance/Behaviour	Other
orofacial movements		new entity	309	174	Other	Appearance/Behaviour
oscillating		new entity	826	558	Other	Other
outraged		new entity	782	524	Appearance/Behaviour	Thought
over-ambitious		new entity	31	27	Other	Other
over-complimentary		new entity	18	15	Appearance/Behaviour	Appearance/Behaviour
over-concerned		new entity	6138	3394	Other	Thought
over-confidence		new entity	177	138	Appearance/Behaviour	Thought
over-dependent		new entity	160	133	Other	Personality
Over-eating	58424009	direct match (finding)	564	372	Appearance/Behaviour	Other
over-happy	85949006	match with new synonym	434	356	Mood/Anxiety/Affect	Mood/Anxiety/Affect
over-inclusive		new entity	82	56	Mood/Anxiety/Affect	Other
Over-optimistic		new entity	104	81	Personality	Other
over-polite		new entity	317	194	Appearance/Behaviour	Appearance/Behaviour

over-protective		new entity	94	63	Other	Other
over-reactive		new entity	98	80	Mood/Anxiety/Affect	Other
over-reliant		new entity	56	44	Other	Personality
over-spending		new entity	115	80	Other	Other
over-stimulated		new entity	294	144	Other	Other
over-stressed	224974006	match with new synonym	503	390	Other	Thought
over-tactile		new entity	232	107	Appearance/Behaviour	Appearance/Behaviour
over-talkative		new entity	1838	1108	Speech	Speech
over-tired	60119000	match with new synonym	334	273	Sleep	Other
overexpansive		new entity	84	74	Mood/Anxiety/Affect	Speech
overfamilarity		new entity	1128	657	Appearance/Behaviour	Appearance/Behaviour
overinvolved		new entity	382	297	Appearance/Behaviour	Other
overly sedated	17971005	direct match (finding)	874	674	Sleep	Appearance/Behaviour
pacing	74691006	direct match (finding)	209300	8930	Appearance/Behaviour	Appearance/Behaviour
paedophilic	84002002	direct match (disorder)	8722	1686	Appearance/Behaviour	Personality
panicy	79823003	direct match (finding)	90	78	Mood/Anxiety/Affect	Mood/Anxiety/Affect
paraesthesia	91019004	direct match (finding)	581	407	Other	Other
paranoia	191667009	direct match (finding)	107828	12155	Thought	Thought
parsimonious		new entity	22	21	Appearance/Behaviour	Other
passive-aggressive	39077006	direct match (finding)	1971	1088	Appearance/Behaviour	Personality
pedantic	286818007	direct match (finding)	390	242	Personality	Personality
peeved		new entity	55	52	Other	Other
pejorative		new entity	153	112	Appearance/Behaviour	Other
pensive		new entity	914	668	Appearance/Behaviour	Appearance/Behaviour

perceived criticism		new entity	214	139	Other	Thought
perceived injustice	216004	match with new synonym	490	287	Other	Thought
perceived insults		new entity	155	89	Other	Thought
perceived persecution	216004	match with new synonym	634	352	Other	Thought
perceived provocation		new entity	988	573	Other	Thought
perceived rejection	105412007	partial match	317	188	Other	Thought
perceived slights		new entity	287	175	Other	Thought
perceived threat	405102006	Direct match (observable entity)	1242	674	Other	Thought
perceived unfairness		new entity	58	35	Other	Thought
perfectionistic	286805004	direct match (finding)	1281	775	Personality	Personality
perfunctory		new entity	138	107	Appearance/Behaviour	Other
perplexed	276245005	direct match (finding)	42838	7234	Appearance/Behaviour	Mood/Anxiety/Affect
persecution	216004	direct match (finding)	119221	12704	Thought	Thought
pessimism	247799003	direct match (finding)	4871	2508	Mood/Anxiety/Affect	Thought
petrified		new entity	727	521	Appearance/Behaviour	Mood/Anxiety/Affect
phlegmatic		new entity	74	61	Appearance/Behaviour	Personality
phobic symptoms		new entity	103	66	Mood/Anxiety/Affect	Other
plaintive		new entity	38	29	Mood/Anxiety/Affect	Other
playing music loudly	247990005	match with new synonym	1480	738	Appearance/Behaviour	Appearance/Behaviour
pliable		new entity	31	21	Appearance/Behaviour	Personality
pompous		new entity	68	38	Appearance/Behaviour	Personality
pondering		new entity	1088	826	Appearance/Behaviour	Other
poor historian	272054000	direct match (finding)	1661	1078	Cognition	Cognition
poor impulse control	286753008	direct match (finding)	1355	682	Appearance/Behaviour	Personality

poorly articulated	286278006	match with new synonym	88	55	Speech	Speech
poorly nourished	248325000	match with new synonym	3082	1530	Appearance/Behaviour	Appearance/Behaviour
posturing	271694000	direct match (finding)	27882	7305	Appearance/Behaviour	Appearance/Behaviour
precocious	235246008	Direct match (qualifier)	99	63	Personality	Other
preconceived ideas		new entity	51	44	Other	Thought
precontemplative		new entity	259	189	Other	Other
prevarication		new entity	218	168	Other	Other
prodromal symtomology		new entity	11	10	Other	Other
prolonged mastication		new entity	19	14	Other	Appearance/Behaviour
promiscuous	85892000	direct match (finding)	1860	833	Appearance/Behaviour	Personality
prosody	35622009	Direct match (observable entity)	1258	758	Speech	Other
pseudo-philosophical		new entity	173	87	Other	Other
pseudoseizures	191714002	direct match (disorder)	336	105	Other	Other
psychomotor agitation	47295007	direct match (finding)	16499	6098	Appearance/Behaviour	Appearance/Behaviour
psychomotor excitation		new entity	816	244	Appearance/Behaviour	Appearance/Behaviour
psychomotor retardation	398991009	direct match (finding)	15038	5051	Appearance/Behaviour	Appearance/Behaviour
psychosocial stress	54427008	partial match	5192	2305	Other	Other
punching walls	477041000000104	match with new synonym	1645	643	Appearance/Behaviour	Other
punctuality		new entity	8226	2719	Other	Personality
querulous		new entity	219	134	Appearance/Behaviour	Other
quite antagonistic		new entity	145	99	Appearance/Behaviour	Appearance/Behaviour
rage	274951009	direct match (finding)	3289	1734	Appearance/Behaviour	Appearance/Behaviour
rambling	162208002	direct match (finding)	9735	3447	Speech	Speech
rapport	710497003	direct match (finding)	195293	16515	Appearance/Behaviour	Appearance/Behaviour

rasping		new entity	57	47	Speech	Speech
ratty		new entity	181	145	Appearance/Behaviour	Mood/Anxiety/Affect
ravenous		new entity	133	104	Appearance/Behaviour	Other
reactionary		new entity	144	115	Appearance/Behaviour	Other
recalcitrant		new entity	54	37	Personality	Other
reclusive		new entity	2326	1202	Appearance/Behaviour	Other
religious interpretations		new entity	53	45	Thought	Other
remorseless		new entity	307	176	Appearance/Behaviour	Other
resentful		new entity	10009	4847	Other	Other
restless	162221009	direct match (finding)	185737	12503	Appearance/Behaviour	Appearance/Behaviour
reticient		new entity	85	58	Appearance/Behaviour	Appearance/Behaviour
rhyming		new entity	1245	599	Speech	Speech
righteous indignation		new entity	13	10	Other	Other
rigidity	311535006	direct match (finding)	16956	5440	Appearance/Behaviour	Appearance/Behaviour
risky behaviour	939841000000102	partial match	11614	3531	Other	Other
ritualistic	67431008	direct match (finding)	5019	1130	Appearance/Behaviour	Appearance/Behaviour
roaming		new entity	918	587	Other	Other
robotic		new entity	653	344	Appearance/Behaviour	Other
rousable		new entity	1439	948	Other	Appearance/Behaviour
rowdy		new entity	268	223	Appearance/Behaviour	Appearance/Behaviour
rude		new entity	43530	7401	Appearance/Behaviour	Appearance/Behaviour
rueful		new entity	115	98	Other	Other
rumination	86110000	direct match (finding)	24435	6444	Thought	Thought
running commentary		new entity	191	110	Thought	Perception

safe guarding issues		new entity	6839	1060	Other	Other
safety behaviours		new entity	1600	689	Mood/Anxiety/Affect	Other
salience		new entity	521	336	Other	Other
sanguine		new entity	157	132	Other	Personality
sardonic		new entity	56	48	Appearance/Behaviour	Other
scantily dressed	248162009	match with new synonym	209	118	Appearance/Behaviour	Appearance/Behaviour
screaming	81916006	direct match (finding)	57753	7238	Speech	Speech
scuffy		new entity	16	15	Appearance/Behaviour	Appearance/Behaviour
secluded	224813006	partial match	5842	2215	Appearance/Behaviour	Other
selective mutism	71959007	direct match (disorder)	395	149	Speech	Speech
self depreciating		new entity	195	119	Other	Thought
self mutilation	130968006	direct match (finding)	449	181	Appearance/Behaviour	Other
self neglect	248054003	direct match (finding)	123614	12649	Appearance/Behaviour	Other
self-recrimination		new entity	63	46	Other	Thought
self-referential thinking		new entity	152	84	Thought	Thought
self-reflective		new entity	1065	430	Other	Other
self-reproachful		new entity	149	101	Other	Mood/Anxiety/Affect
selfishness		new entity	2060	1335	Personality	Other
semantic errors		new entity	39	20	Cognition	Speech
sexual disinhibition	225533009	direct match (finding)	36661	3789	Appearance/Behaviour	Appearance/Behaviour
sexual inappropriateness	248099006	match with new synonym	40919	4243	Appearance/Behaviour	Appearance/Behaviour
sexually promiscuous	85892000	direct match (finding)	527	245	Appearance/Behaviour	Personality
shadow boxing		new entity	3483	501	Appearance/Behaviour	Other
short tempered		new entity	1797	1145	Appearance/Behaviour	Mood/Anxiety/Affect

shrieking		new entity	266	167	Speech	Speech
silently mouthing		new entity	40	25	Speech	Speech
slight unsteadiness	271713000	direct match (finding)	32	28	Appearance/Behaviour	Appearance/Behaviour
slightly vacant	285741001	direct match (finding)	132	109	Appearance/Behaviour	Mood/Anxiety/Affect
smashed computers	284661005	match with new synonym	65	34	Appearance/Behaviour	Other
smashing property	284661005	match with new synonym	11	10	Appearance/Behaviour	Other
smashing tv	284661005	match with new synonym	227	102	Appearance/Behaviour	Other
snappy		new entity	5009	2727	Appearance/Behaviour	Mood/Anxiety/Affect
snobbishness		new entity	62	50	Other	Other
snorting noises		new entity	26	12	Speech	Speech
sobbing	28263002	direct match (finding)	1688	934	Appearance/Behaviour	Mood/Anxiety/Affect
socially phobic	25501002	direct match (finding)	95	67	Mood/Anxiety/Affect	Other
soliloquy		new entity	108	67	Speech	Speech
solitary		new entity	2804	1495	Appearance/Behaviour	Other
somatic complaints	386738004	direct match (finding)	1474	614	Other	Mood/Anxiety/Affect
somatic hallucinations		new entity	4345	1587	Perception	Perception
somatic passivity	281144009	direct match (finding)	2692	1374	Thought	Thought
somatic preoccupations		new entity	215	135	Other	Thought
somatic sensations	397725003	direct match (finding)	389	224	Perception	Perception
somatisation	397923000	direct match (disorder)	1084	498	Mood/Anxiety/Affect	Mood/Anxiety/Affect
sombre		new entity	1108	725	Mood/Anxiety/Affect	Other
somewhat overdressed	248162009	match with new synonym	41	33	Appearance/Behaviour	Appearance/Behaviour
somnolent		new entity	160	107	Appearance/Behaviour	Other
soporific		new entity	76	63	Sleep	Other

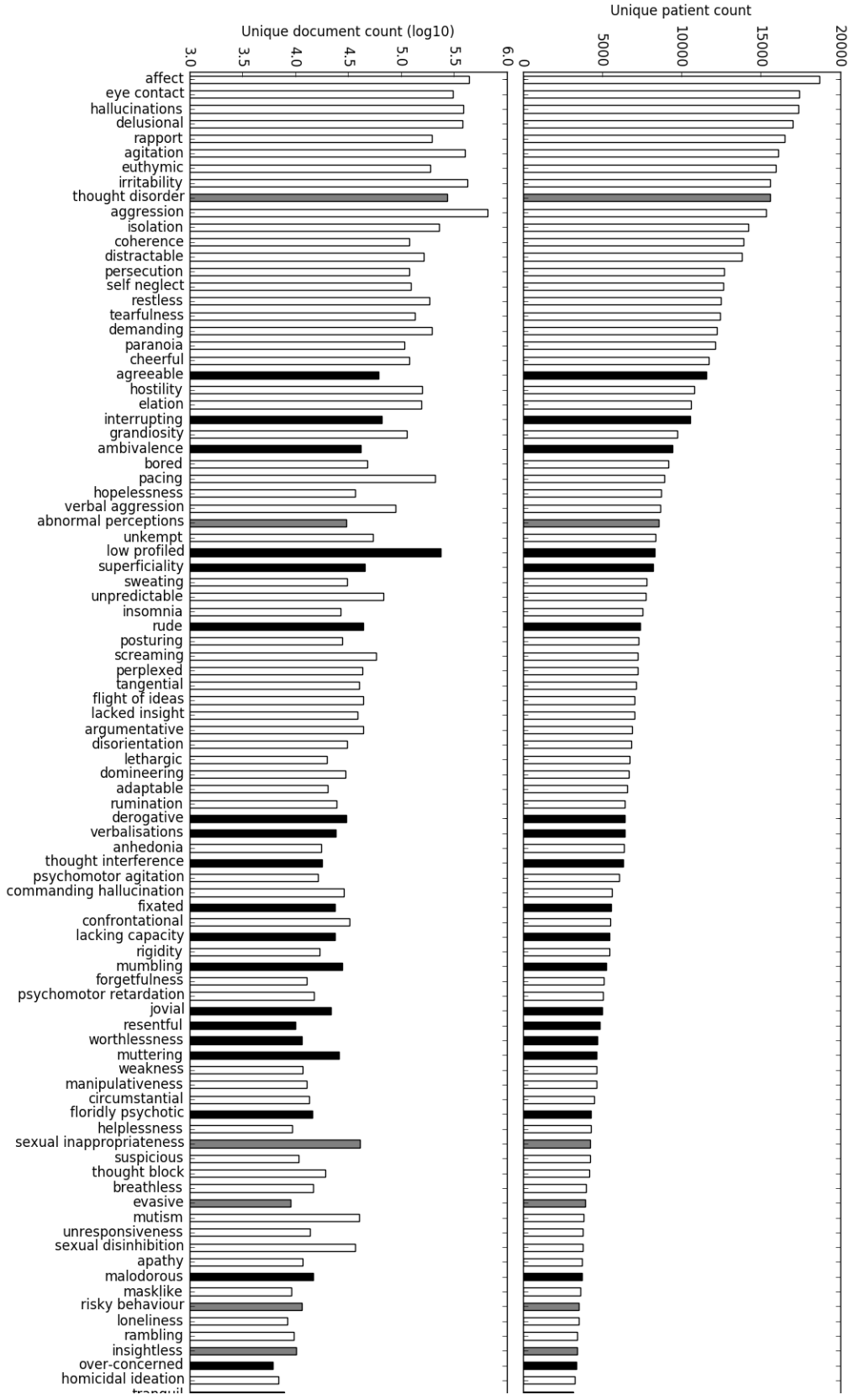
spending recklessly		new entity	107	69	Appearance/Behaviour	Other
spiritual quest		new entity	35	22	Other	Other
squalid conditions	224271009	direct match (finding)	586	266	Other	Other
staccato quality	102936006	match with new synonym	18	11	Speech	Speech
stammering	39423001	direct match (finding)	3304	696	Speech	Speech
stare menacingly		new entity	18	11	Appearance/Behaviour	Mood/Anxiety/Affect
stared intimidatingly		new entity	15	13	Appearance/Behaviour	Appearance/Behaviour
stereotyped	84328007	direct match (finding)	3281	1530	Appearance/Behaviour	Appearance/Behaviour
stern		new entity	5751	2018	Appearance/Behaviour	Other
stigmatizing		new entity	3754	2191	Other	Other
stoic		new entity	425	310	Appearance/Behaviour	Personality
stroppy		new entity	511	416	Appearance/Behaviour	Mood/Anxiety/Affect
stubborn	286806003	direct match (finding)	2366	1403	Personality	Personality
stupor	89458003	direct match (finding)	1021	452	Appearance/Behaviour	Appearance/Behaviour
sulk		new entity	551	404	Appearance/Behaviour	Other
supercilious		new entity	96	53	Personality	Personality
superficial brightness		new entity	1226	563	Mood/Anxiety/Affect	Other
superficial relationships		new entity	123	88	Personality	Personality
superficiality		new entity	45375	8224	Appearance/Behaviour	Appearance/Behaviour
suspicious	22927000	direct match (finding)	10760	4226	Thought	Thought
swaggering		new entity	113	68	Appearance/Behaviour	Appearance/Behaviour
sweating	415690000	direct match (finding)	31049	7815	Other	Appearance/Behaviour
systematised delusions	40277007	direct match (finding)	918	470	Thought	Thought
taciturn		new entity	345	238	Speech	Appearance/Behaviour

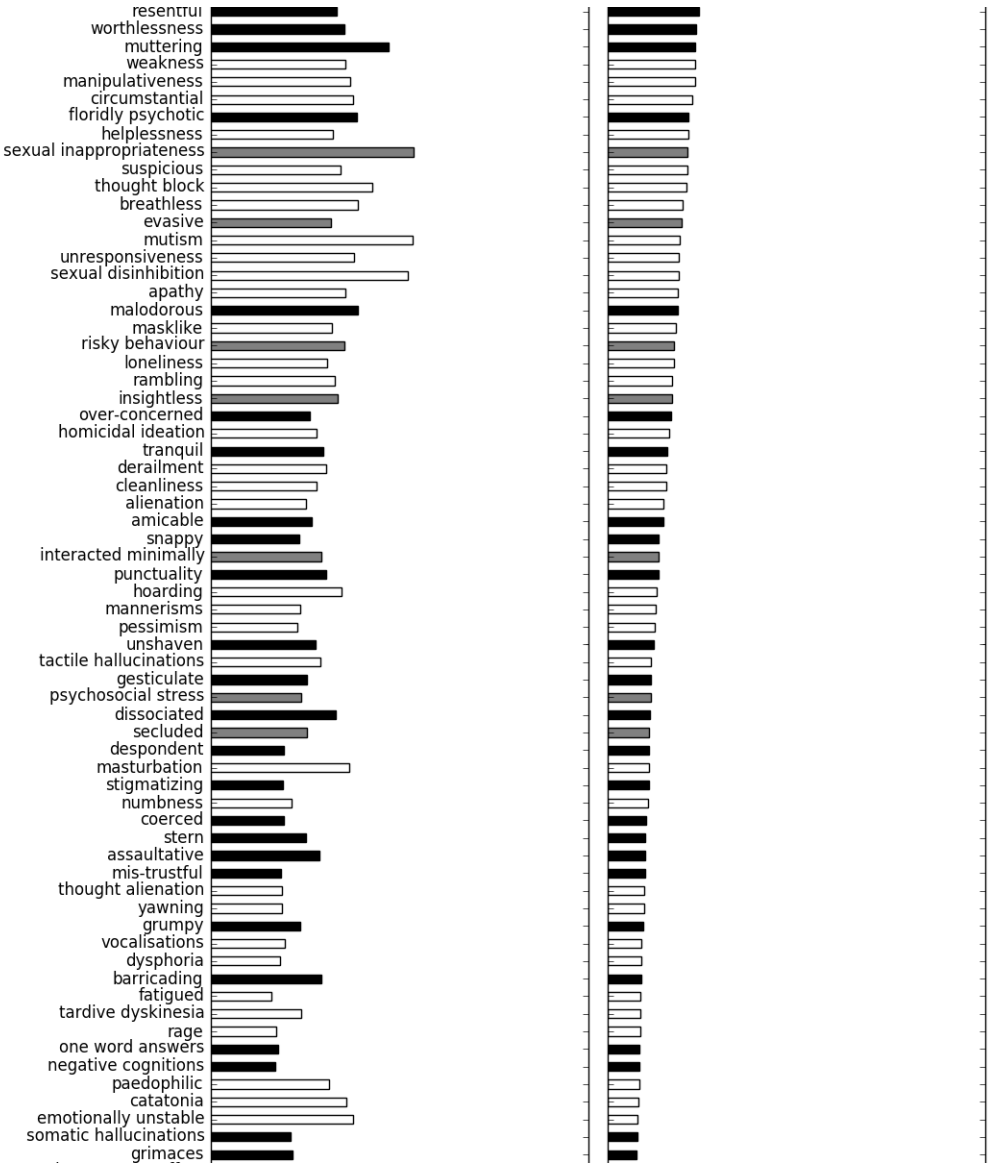
tactile hallucinations	66609003	direct match (finding)	7442	2329	Perception	Perception
talk excessively		new entity	1404	702	Speech	Speech
talkative		new entity	107	96	Speech	Speech
talking incessantly		new entity	885	543	Speech	Speech
talking nonsense		new entity	1114	668	Speech	Speech
tangential	74396008	direct match (finding)	39841	7163	Speech	Speech
tardive dyskinesia	102449007	direct match (disorder)	5260	1748	Other	Appearance/Behaviour
tearfulness	271951008	direct match (finding)	135719	12432	Appearance/Behaviour	Mood/Anxiety/Affect
telegraphic		new entity	420	218	Other	Speech
testy		new entity	34	33	Other	Mood/Anxiety/Affect
tetchiness		new entity	141	118	Appearance/Behaviour	Personality
theatrical		new entity	520	331	Appearance/Behaviour	Other
thought alienation	247695002	direct match (finding)	3710	1936	Thought	Thought
thought block	2899008	direct match (finding)	19027	4195	Speech	Speech
thought disorder	41591006	match with new synonym	270817	15570	Speech	Speech
thought interference		new entity	18014	6319	Thought	Thought
tonic clonic (seizures)	54200006	direct match (finding)	1190	399	Other	Other
tranquil		new entity	7794	3189	Other	Other
transference	224982006	direct match (finding)	978	415	Other	Other
transfixed		new entity	186	140	Appearance/Behaviour	Other
truancy	105479008	direct match (finding)	4206	1126	Other	Other
truculent		new entity	262	183	Appearance/Behaviour	Personality
tuberosity		new entity	110	67	Other	Other
tumultuous		new entity	179	112	Other	Other

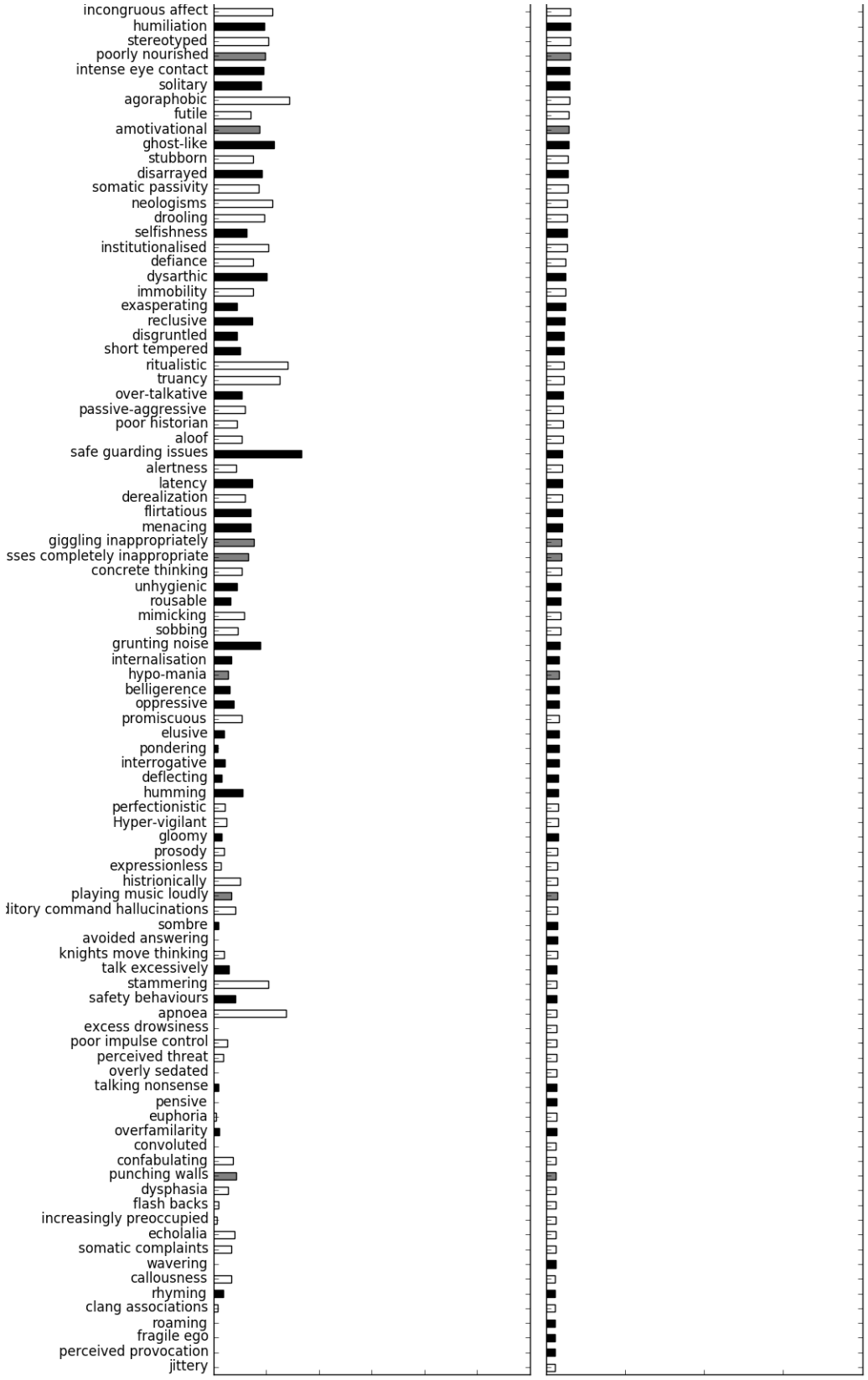
under-dressed	248162009	match with new synonym	37	28	Appearance/Behaviour	Appearance/Behaviour
unhygienic		new entity	1666	951	Appearance/Behaviour	Other
unintelligible noises		new entity	48	25	Speech	Speech
uninteractive		new entity	126	94	Appearance/Behaviour	Appearance/Behaviour
unkempt	46017004	direct match (finding)	54033	8378	Appearance/Behaviour	Appearance/Behaviour
unpredictable	284509004/225656007	direct match (finding)	68039	7733	Other	Other
unrealistic ideas	247785002	partial match	654	455	Other	Thought
unrelenting standards		new entity	104	54	Other	Other
unresponsiveness	422768004	direct match (finding)	13630	3796	Speech	Appearance/Behaviour
unshaven		new entity	6754	2478	Appearance/Behaviour	Appearance/Behaviour
verbal aggression	248003003	direct match (finding)	87934	8663	Speech	Appearance/Behaviour
verbalisations		new entity	24315	6410	Speech	Speech
vexed		new entity	337	232	Other	Other
vindictive		new entity	413	289	Personality	Personality
violent fantasies		new entity	6404	462	Thought	Thought
virtually uninterruptible		new entity	21	18	Speech	Speech
vivacious		new entity	118	94	Appearance/Behaviour	Mood/Anxiety/Affect
vocalisations	278288005	direct match (finding)	3891	1825	Speech	Speech
vociferous		new entity	493	349	Personality	Other
wasting such money	300692004	match with new synonym	245	206	Appearance/Behaviour	Other
wavering		new entity	911	608	Other	Other
waxy	13052006	direct match (finding)	605	306	Appearance/Behaviour	Appearance/Behaviour
weakness	13791008	direct match (finding)	11762	4659	Other	Other
whispery	21313003	match with new synonym	26	19	Speech	Speech

wizardry		new entity	62	23	Other	Other
wondersome		new entity	38	19	Other	Other
worthlessness		new entity	11607	4721	Mood/Anxiety/Affect	Mood/Anxiety/Affect
wronged		new entity	664	445	Other	Thought
yawning	248626009	direct match (finding)	3668	1933	Appearance/Behaviour	Appearance/Behaviour
yelping noises		new entity	21	12	Speech	Speech

5.1.2 Supplementary File 2


Unique document count (log10)
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Unique patient count
0
5000
10000
15000
20000
affect
eye contact
hallucinations
delusional
rapport
agitation
euthymic
irritability
thought disorder
aggression
isolation
coherence
distractable
persecution
self neglect
restless
tearfulness
demanding
paranoia
cheerful
agreeable
hostility
elation
interrupting
grandiosity
ambivalence
bored
pacing
hopelessness
verbal aggression
abnormal perceptions
unkempt
low profiled
superficiality
sweating
unpredictable
insomnia
rude
posturing
screaming
perplexed
tangential
flight of ideas
lacked insight
argumentative
disorientation
lethargic
domineering
adaptable
rumination
derogative
verbalisations
anhedonia
thought interference
psychomotor agitation
commanding hallucination
fixated
confrontational
lacking capacity
rigidity
mumbling
forgetfulness
psychomotor retardation
jovial
resentful
worthlessness
muttering
weakness
manipulativeness
circumstantial
floridly psychotic
helplessness
sexual inappropriateness
suspicious
thought block
breathless
evasive
mutism
unresponsiveness
sexual disinhibition
apathy
malodorous
masklike
risky behaviour
loneliness
rambling
insightless
over-concerned
homicidal ideation


resentful
worthlessness
muttering
weakness
manipulativeness
circumstantial
floridly psychotic
helplessness
sexual inappropriateness
suspicious
thought block
breathless
evasive
mutism
unresponsiveness
sexual disinhibition
apathy
malodorous
masklike
risky behaviour
loneliness
rambling
insightless
over-concerned
homicidal ideation
tranquil
derailment
cleanliness
alienation
amicable
snappy
interacted minimally
punctuality
hoarding
mannerisms
pessimism
unshaven
tactile hallucinations
gesticulate
psychosocial stress
dissociated
secluded
despondent
masturbation
stigmatizing
numbness
coerced
stern
assaultative
mis-trustful
thought alienation
yawning
grumpy
vocalisations
dysphoria
barricading
fatigued
tardive dyskinesia
rage
one word answers
negative cognitions
paedophilic
catatonia
emotionally unstable
somatic hallucinations
grimaces


incongruous affect
humiliation
stereotyped
poorly nourished
intense eye contact
solitary
agoraphobic
futile
amotivational
ghost-like
stubborn
disarrayed
somatic passivity
neologisms
drooling
selfishness
institutionalised
defiance
dysarthic
immobility
exasperating
reclusive
disgruntled
short tempered
ritualistic
truancy
over-talkative
passive-aggressive
poor historian
aloof
safe guarding issues
alertness
latency
derealization
flirtatious
menacing
giggling inappropriately
sses completely inappropriate
concrete thinking
unhygienic
rousable
mimicking
sobbing
grunting noise
internalisation
hypo-mania
belligerence
oppressive
promiscuous
elusive
pondering
interrogative
deflecting
humming
perfectionistic
Hyper-vigilant
gloomy
prosody
expressionless
histrionically
playing music loudly
litory command hallucinations
sombre
avoided answering
knights move thinking
talk excessively
stammering
safety behaviours
apnoea
excess drowsiness
poor impulse control
perceived threat
overly sedated
talking nonsense
pensive
euphoria
overfamilarity
convoluted
confabulating
punching walls
dysphasia
flash backs
increasingly preoccupied
echolalia
somatic complaints
wavering
callousness
rhyming
clang associations
roaming
fragile ego
perceived provocation
jittery

5.1.3 Errata

The original manuscript states *"As we were primarily intending to identify initial clusters for validation by clinical experts it was felt that single epoch of training, over the 20M clinical records available, was sufficient."*

This statement is incorrect. The actual count of documents used for training was 11 745 094 as stated elsewhere.

5.1.4 Discussion

In the above manuscript, I make the claim *"Given a sufficiently large corpus of documents, typically written by hundreds of clinical staff over several years, it is often difficult to track the evolution of vocabulary used within the local EHR setting to describe potentially important clinical constructs."*. The language of psychiatry is necessarily diverse, owing to the fact that the primary diagnostic instrument for large swathes of the field is via the communication patients have with their doctor. This introduces a high degree of subjectivity into the field, as patient's are unlikely to be able to express their experience of symptoms in the preferred formulisations of the medical profession. Rather, symptomatology (especially in the case of SMI), is likely to be expressed in colloquialisms. This phenomena has been studied wth regard to the ICD-9 disease classification system [239]. Here, Forbush et al found that, from a sample of 750 clinical notes, only 36% of symptom expressions are represented in ICD-9. As it seems, colloquial phrases that manage to find their way into the EHR are presumably subject to all of the neologisms, neoidioms and other phenomena that is observable on general English.

The MedLex project [188] sought to build a clinical lexicon, to facilitate the mapping of 'real world' usage of clinical text to describe symptoms to standard terminologies, for instance, by taking into account morphosyntactic and semantic information. By analysing a corpus of 55 million clinical notes, the authors managed to map approximately 40% of the tokens to UMLS entities (the authors account for the rest as being likely misspellings or proper nouns). In this work, the authors base their work on attempting to map clinical language to known entities within clinical terminologies. Further work might integrate their approach with my efforts described above in identifying novel SMI symptomatology. For instance. the MedLex software could be used to map existing symptom entities to terminologies, negating the need for this to be done manually. However, MedLex does not offer the means to identify and explore concepts outside of the UMLS terminologies, as it does not offer any kind of relatedness metric for un-mapped terms. For this reason, the two approaches can be synergistically combined.

Further improvements to the process I describe in the above work might include automation of certain elements, such as implementing a spell checking and or morphological normalisation step to reduce the about of manual curation required. Although such components would be trivial to implement, I regarded them with caution - as the CRIS corpus has not been studied in such a fashion previously, any additional heuristics would need to be evaluated with regard to their impact on my stated goal. Relative to the amount of manual curation required (which amounted to checking a list several thousand n-grams), implementing such heuristics did not seem to warrant the risk of creating a systemic bias.

In order to produce the seed concept list necessary for the k-means clustering algorithm to identify related n-grams, is was necessary to curate a list of unambiguous mental health symptom n-grams. This is due to a limitation of the Word2Vec algorithm, in that it is not capable of handling different 'sense' interpretations of the same word. Therefore, words like 'associations' from 'loosening of associations', are not likely to project near relevant symptomatology terms, since they can be used in many contexts (such as 'housing associations'). Similarly, 'flight of ideas' also contains words that might be difficult to disambiguate. However, a common abbreviation used by psychiatrists for this latter symptom is 'foi', which enabled the inclusion of this symptom. The specific terms used for identifying relevant clusters are listed in the right column of table 1 in the manuscript. Notably, this limitation of Word2Vec has since been addressed by the use of contextualised word embeddings [240], and future work should explore such techniques for better word representations.

One notable result from the above work was the relatively low Kappa agreement score of 0.45 for the curated concepts. Although this signals a substantial degree of different interpretation of the extracted concepts, it is perhaps not unexpected due to 1) the explorative nature of the task and 2) the known difficulty of achieving reasonable Kappa scores on clinical data, even for well defined tasks [241, 242]. Nevertheless, the implications of the low Kappa highlight the need for clinical research and discussion of novel symptom observations before work is undertaken to have them formally accepted into standard terminologies.

5.1.5 Conclusion

Amongst the various medical domains, mental health is particularly notable for its heavy use of free text to describe clinically relevant observations [1, 64]. This presents certain challenges for phenotyping SMI from EHR data, as SMI symptom entities as depicted in SNOMED CT are often encoded within the free text, as opposed to structured inputs. IE seeks to represent unstructured text in a structured format. Approaches in this context

often make use of curated or semi-curated terminology resources such as SNOMED CT or the broader Unified Medical Language System [178, 183, 243]. Such resources can help in the construction of lexical resources of medical terminology and relevant synonymy, which are useful for mapping text to resources. However, as they were never designed for specific NLP use cases, it is widely acknowledged that such dictionary based methods struggle to capture the vast diversity of how clinical constructs can represented in natural language [244]. Although such resources often include provisions to address synonym usage, it is unlikely that they will ever offer complete coverage for the vast lexicon offered by the English language. Here, word embedding approaches may assist when adapting existing IE paradigms to new corpora of clinical text by suggesting the most salient synonyms and novel entities within the corpus.

Referring back to my motivation for this thesis, the ultimate goal of this work was to make the described methodology available for broader exploitation by non-technical clinical researchers. However, due to time constraints this was not possible and the model remains accessible only via an API. Nevertheless, I hope to have demonstrated a proof of principal for the approach. In addition, I have produced evidence to prompt further investigation into extending our knowledge of the clinical relevance of real world SMI symptom constructs by looking beyond the favoured terminology of the NHS and examining clinical text directly.

In the next chapter, I shift focus from fundamental definitions of symptomatology and their extraction, to broader elements of industrialising IE processes within the confines of the NHS environment.

Chapter 6

CogStack - Enterprise Architecture for Information Retrieval and Extraction in Resource Constrained Environments

6.1 Overview

Defining an IE problem and creating solutions is one aspect of the goals set out in chapter 3. Applying the techniques at a scale to meet the demands of enterprise analytics is a rather different problem. Models for processes such as IE must not to be 'static', in the sense that in order for them to remain accurate over time, they need to be periodically re-trained and re-validated on new data as it is produced. Once updated, they need to be re-deployed over all available data in order to yield the most accurate predictions. When dealing with hundreds of millions of clinical documents, this feedback loop requires that a substantial amount of compute be made available such that data can be provisioned in a timely fashion.

Grid frameworks and cloud computing platforms such as the Kubernetes Engine, SLURM, the Hadoop ecosystem, Google Cloud Dataflow and Microsoft Azure Databricks create the environment for elastic resourcing, to allow high throughput of so called 'embarassingly parrallel' problems such as IE. Despite a 2013 Government mandate for public cloud adoption, barriers ranging from financial, technical and/or lack of clarity over information governance reasons have meant many NHS Trusts have been slow to move their

infrastructure onto such platforms. In turn, even well resourced Trusts suffer from an acute lack of NLP and other large scale analytics capability, representing an opportunity loss via their inability to make use of their administrative data for business intelligence, research and other use cases. In this chapter, I move away somewhat from the specific focus of NLP in the domain of SMI symptomatology, and take a broader view of the opportunities and challenges of clinical NLP within the NHS.

6.2 Overcoming Scaling Issues in Model Deployment

Having derived models for SMI symptomatology via TextHunter (as described in previous chapters), it became apparent that a framework was missing to optimise the deployment of such models. That is to say, although TextHunter is capable of scaling vertically via multi-threading, this method of scaling is limited to a single server. This is also true of the GATE framework on which TextHunter is built, in the sense that it does not have native support for operating in a distributed fashion. The maximum throughput by which large amounts of data can be processed is therefore limited to the number of cores that can be provisioned on a single server. Under circumstances where one wishes to deploy tens or even hundreds of NLP processes simultaneously across large volumes of clinical documents on a regular basis, such an architecture is impractical - even with a highly optimised codebase. In order to use the SMI symptom models (or indeed, any NLP pipeline) to meet the day to day processing requirements of a typical NHS Trust, a distributed processing architecture is required.

In addition, the evolution of a range of open source NLP technologies during the course of my PhD enabled further analytics opportunities based on clinical text. Improvements to the Apache Tika project (binary file format conversion), Tesseract (Optical Character recognition), Elasticsearch and Solr (Information retrieval) opened new doors in how one might address fundamental architecture questions about how an Extract, Transform, Load (ETL) process might be built in order to create an analytics environment.

An opportunity arose to address such a scenario at King's College Hospital Foundation Trust, via funding from the 100 000 Genome Project [245]. Given the lack of grid and cloud resources, the principal challenge was to create an architecture capable of identifying candidate patients for recruitment into the project, and subsequently extracting their phenotypic data from both structured and unstructured parts of the clinical record.

The below paper describes the work in building this architecture, as well as general findings about the opportunities and challenges of informatics within the NHS (author contributions are listed in the paper).

Jackson *et al. BMC Medical Informatics and Decision Making* (2018) 18:47
https://doi.org/10.1186/s12911-018-0623-9

BMC Medical Informatics and Decision Making

SOFTWARE Open Access

CogStack - experiences of deploying integrated information retrieval and extraction services in a large National Health Service Foundation Trust hospital

Richard Jackson[1,2*†], Ismail Kartoglu[7†], Clive Stringer[3], Genevieve Gorrell[4], Angus Roberts[4], Xingyi Song[4], Honghan Wu[1,8], Asha Agrawal[3], Kenneth Lui[5], Tudor Groza[6], Damian Lewsley[3], Doug Northwood[3], Amos Folarin[1,5], Robert Stewart[1,2] and Richard Dobson[1,5]

Abstract

Background: Traditional health information systems are generally devised to support clinical data collection at the point of care. However, as the significance of the modern information economy expands in scope and permeates the healthcare domain, there is an increasing urgency for healthcare organisations to offer information systems that address the expectations of clinicians, researchers and the business intelligence community alike. Amongst other emergent requirements, the principal unmet need might be defined as the 3R principle (right data, right place, right time) to address deficiencies in organisational data flow while retaining the strict information governance policies that apply within the UK National Health Service (NHS). Here, we describe our work on creating and deploying a low cost structured and unstructured information retrieval and extraction architecture within King's College Hospital, the management of governance concerns and the associated use cases and cost saving opportunities that such components present.

Results: To date, our CogStack architecture has processed over 300 million lines of clinical data, making it available for internal service improvement projects at King's College London. On generated data designed to simulate real world clinical text, our de-identification algorithm achieved up to 94% precision and up to 96% recall.

Conclusion: We describe a toolkit which we feel is of huge value to the UK (and beyond) healthcare community. It is the only open source, easily deployable solution designed for the UK healthcare environment, in a landscape populated by expensive proprietary systems. Solutions such as these provide a crucial foundation for the genomic revolution in medicine.

Keywords: Elasticsearch, Electronic health records, Information extraction, Clinical informatics, Natural language processing

*Correspondence: richgjackson@gmail.com
†Richard Jackson and Ismail Kartoglu contributed equally to this work.
[1]Institute of Psychiatry, Psychology and Neuroscience, King's College London, 16 De Crespigne Park, SE5 8AF London, UK
[2]South London and Maudsley NHS Foundation Trust, Denmark Hill, SE5 8AZ London, UK
Full list of author information is available at the end of the article

Background

Large healthcare organisations are often responsible for provisioning care in a wide range of medical specialties. It is not uncommon for a given speciality to make use of bespoke IT systems to support the specific requirements of clinicians at the point of care, such as imaging technologies, electronic prescribing and intensive care monitoring. This leads to a tendency for healthcare IT departments to support a large number of systems, which often suffer from integration issues, in the sense that there may not be a single interface that allows users to access data across all systems simultaneously. While there have been many attempts to standardise intra-system communication with the use of controlled languages and data schemas, such as HL7 [1], the myriad of vendors, differential versioning of the standards and the ambiguity in the interpretation of the standards has caused such efforts to be only partially successful in practice[2–4]. This has lead to a high degree of heterogeneity in how information is managed within and between different NHS Trusts, which in turn has inflated the costs of creating suitable data management and analytics solutions, due to the investment required for successful implementation. For the end user, whether they be a clinician, a researcher or a business intelligence analyst, the implication is often described as a 'needle in a haystack' problem, owing to the complexity of how, where and why data is stored in a host of disparate sources. Without significant guidance from central hospital IT departments, many lay users of health information systems may not be aware of the logic of how data flows between them, and thus opportunities to use the organisation's data to drive efficiency improvements are undermined.

The problem is further compounded by the nature of health data. In contrast to domains where structured data are captured in abundance (for example in e-commerce customer behaviour, retail loyalty card usage and financial trading patterns), all but a thin supernatant of clinical information are recorded as unstructured data in the form of the clinical narrative, via free text clinical notes, discharge summaries and referral letters [5, 6]. Since unstructured data are inherently more difficult to manage and query, this preference of clinicians manifests as a complication in how data can be provisioned between stakeholders effectively.

Information retrieval technologies have the stated aim of providing the ability to filter very large quantities of both structured and unstructured information and return relevant results at high speed. Due to their relatively straight-forward manner of ingesting data without a requirement to pre-define a schema, they have enjoyed a long history of success in almost every domain of information management, and are deployed in business critical environments such as enterprise document retrieval, bioinformatics, e-commerce and log management. Typically, they are provisioned through a simple, intuitive interface by which a user can query structured and unstructured data simultaneously, and rapidly refine their query to provide results relevant to their intent. This feature of query refinement through iteration is especially important in healthcare, given the nature of the medical 'sub-language', where concepts tend to be represented in clinical text with a high degree of assumed knowledge and a low level of verbosity [7, 8].

When correctly implemented in a healthcare organisation, such technologies are increasingly employed to overcome a range of data accessibility issues. We delineate these issues by what we refer to as the 3R principle:

Right data With large amounts of data flowing through an organisation, often conflicting reports may occur. For example, two different diagnoses may be reported on two separate occasions. A third party who only has access to a partial view of the data would not be able to make a judgement on the current status of a particular patient. Therefore, maximising the recall (sensitivity) of an information retrieval system is essential to ensure data sufficiency for a question answering system. On the other hand, almost a decade of widespread EHR adoption has created a deluge of data in many progressive healthcare organisations - a trend which is certain to grow. A key consideration reflecting the usability of an information retrieval system is therefore also its ability to avoid false positives (precision, or positive predictive value) and not overburden the user with irrelevant results.

Right place Many enterprise grade approaches to integration opt for data-warehousing methods to provide a single end point, often a SQL relational database, to offer an online analytical processing (OLAP) style capability. While the value of such approaches is well established, it is often restricted to users of the business intelligence community, and generally limited in its ability to effectively manage free text. This constraint therefore inhibits users elsewhere in the organisation, who may have simpler requirements regarding data use (for example, to find documents relating to patients in their care that contain certain keywords). In addition, the technical skills required to use OLAP resources effectively may concentrate in a relatively small number of individuals. Therefore, user-friendly solutions with a lower technical barrier for effective use will enable a degree of 'self provisioned analytics' and thus enjoy a wider uptake amongst employees.

Right time Time based factors are often the difference between actionable and 'stale' information in clinical and business decision making. For instance, the opportunity to code clinical documents for repatriation may be lost if

relevant data cannot be supplied to a code billing team within a commercial deadline. Similarly, if the data deluge negate the possibility of a human reading every document, there is potential to under-code the dispensation of high cost drugs and/or services. In the case of critical care, identifying antagonistic factors towards recovery at speed may help to deliver more favourable outcomes. The requirement to make data available throughout the organisation with as little latency as possible is critical to ensure its effective use.

Information governance

The aim of our project is to offer a general information retrieval system and OLAP analytics capability to meet the requirements of a large variety of use cases. However, in order to protect the rights of individuals as per the UK 1998 Data Protection Act, there are strict controls on how different types of data can be used for different purposes. From a technical perspective, this imposes limitations on how and where data can be provisioned and what transformations it must undergo. Generally speaking, the individuals within a given dataset may be classed as identifiable (no information is removed), pseudonymised (identifiers replaced by a pseudonym, enabling data linkage to other datasets), or anonymised (all identifiers removed, or data aggregated such that re-identification of individuals is nearly impossible). Each class of information removal represents different levels of risk regarding the secondary use of data. Although the details of the Act are complex, the practical applications in a clinical setting might be summarised in the following scenarios:

Business intelligence Activities that utilise the data a Trust holds for the purposes of improving its operational efficiency. Here, named functions within the Trust may use identifiable data for a limited number of well defined purposes. For example, the Trusts clinical coding function has the remit to examine data generated in the course of a patient's care, to ensure that delivered clinical services are accurately recorded and billed for. Alternatively, the Trust may use its data to meet its legal requirements to report figures to central government departments concerning the organisation's performance or indicators of the nation's health, such as cancer survival or diabetes rates.

Service improvement activities Under approval from the Trust's appointed Caldicott guardian, Trust staff may access pseudonymised data in limited amounts in order to undertake internal research projects with the aim of improving the quality and/or efficiency by which a Trust delivers clinical services. The criteria for this generally requires that the affected patients will potentially directly benefit from the project outcomes. For example, this scenario might be invoked if a clinician is seeking to challenge current practices in service delivery, such as how the length of inpatient of hospital visits are predicted in order to reduce the number of staff hours invested in this task.

Enclave style research environment An increasingly common method by which non-staff researchers are able to access clinical record data. Similar to service improvement activities, this method covers an expanded scope that enables clinical data to be used for research projects beyond direct patient benefit. Here, external parties may access pseudonymised and de-identified clinical data in limited quantities in highly secure environments under ethical agreements granted by UK Research Ethics Councils. Examples include Clinical Records Interactive Search [9] and Secure Anonymised Information Linkage Databank [10].

Explicitly obtained consent Perhaps the most common method of accessing clinical data for research is by explicitly obtaining consent from patients to use their identifiable data. This is also governed by Research Ethic Councils, and generally involves strict practices to guard against data breach. Although the most liberal in terms of how the data can be used (since patients are directly briefed as to the nature of the research and how their details will be used), the resource intensive means by which consent must be obtained generally creates a practical limit on the number of patients that can be included in such studies. In turn, this affects the type of study for which this approach is suitable.

Implementation

Here, we describe our work on the CogStack architecture, an open source information retrieval and extraction architecture to provide an alternative to the UK healthcare community in a space traditionally occupied by commercial vendors. We describe its features and how it has been implemented within King's College Hospital (KCH). We focus specifically on surfacing the deep data with the EHR for identification and recruitment of patients into the 100k Genomics England Project [11], for which the concept was funded and developed via NHS England Enablement Funding. Finally, we explore a vision of how such technology can be exploited for a range of use cases within the modern hospital environment.

Previous work

There are several reports of systems that offer information retrieval solutions directed at the challenges within the healthcare domain. Moen et al. [12] proposed a variety for methods for selecting similar care episodes from other

patients, given a particular case of interest. The NLP-Pier concept [13] combines an information retrieval and a biomedical entity information extraction system based around the popular open source project Elasticsearch. In the UK, comparable projects that make use of information retrieval systems include the CRIS [14] project, which uses the commercial FAST search engine and a custom text de-identification algorithm to make clinical notes from mental health patients available for research.

Cognition

In addition, the open source Cognition platform [15, 16] is a vertically and horizontally scalable application that retrieves binary encoded documents and plain text from a relational database, and optionally de-identifies personal identifiers (for example, patient names, addresses and phone numbers) in text.

During routine clinical administrative activity, PHIs are often routinely collected as semi-structured data during the course of a patients care (for example, patient and carer names, addresses, NHS numbers and dates of birth). Such information is a valuable source of data for de-identification methods, as it offers highly precise information about the nature of the text strings that should be removed. However, in natural language, PHIs are often written in a variety of formats, requiring that high accuracy approaches have a greater flexibility that can be achieved by simple direct string matching. For instance, an address written "Institute of Psychiatry, Psychology and Neuroscience, 16 De Crespigne Park, SE5 8AF" might be shortened to "Institute of Psychiatry, 16 De Crespigne Park SE5 8AF". Similarly, PHIs in natural language documents may contain spelling mistakes or additional punctuation tokens. To achieve flexibility, the effectiveness of rules based approaches has been demonstrated elsewhere [17]. The Cognition de-identification algorithms, which are used in this work, are designed to take into account misspellings, tolerance for missing/redundant information, and word order without the need for manual rule crafting nor construction of labelled datasets for machine learning approaches, which are known to be an expensive process [18]. Cognition applies a "sliding window" approach to detect the regions of text where patient identifiers are mentioned. During the processing of a document, the patient specific PHIs are retrieved from semi-structured fields in a database, and the Levenstein edit distance is calculated for each PHI token at every character offset available in the document. If the Levenstein distance is above a configurable threshold, the offsets of the match are masked. This allows for an efficient method of removing PHIs in a document, even if they are misspelt in the document or source inputs.

The de-identified output text from Cognition contains meta-data related to patients and the document such as a hash code of a combination of the patient's identifiers and document date, which are useful for version control. The output text may also be output to a relational database or Elasticsearch index, to be used by downstream services such as the Kibana web interface, or natural language processing applications. Cognition uses the Apache Tika library for converting common document formats such as Microsoft Word, PDFs, Excel etc. into text and further applies Optical Character Recognition to scanned documents that are only available in image formats (including scanned PDFs) using the Tesseract library. Cognition handles horizontal scaling by using a HTTP-based coordinator-client approach where a coordinator assigns work coordinates to the clients.

The CogStack architecture

CogStack is a set of open source and open core services, co-ordinated by a batch processing framework that builds on the concepts of the Cognition platform by offering additional interfaces for NHS systems and NLP technologies. Out-of-the-box open source components were selected from a variety of successful open source and freely licensed projects. The services can be deployed using the Docker containerisation technology, to maximise ease of deployment.

The overall goal of the architecture is to undertake a series of configurable transformations of clinical data housed in relational databases and to load the transformed data into an Elasticsearch information retrieval engine (otherwise known as a search engine - described below), whereupon the 3R principles can be more readily addressed than via direct communication with the untransformed source databases alone. Each transformation is highly configurable, in accordance with the desired use case of the end product. For example, it is not necessary (or even desirable) to de-identify data for business intelligence use cases, and thus this can be disabled. Similarly, not all use cases will require computationally expensive entity extraction NLP processes. The rationale for the choice of components is described below, while the flow of data and transformations in the CogStack architecture is described in Fig. 1.

Handling text and other unstructured data

During a patient's course of treatment, a large number of documents such as referral letters and discharge summaries tend to be generated via word processing applications, predominently Microsoft Word. In addition, such documents may undergo further manipulations, such as PDF conversion and printing and rescanning as an image before they reach their final storage location, usually a relational database. Such manipulations represent complications for search and NLP applications, as the valuable

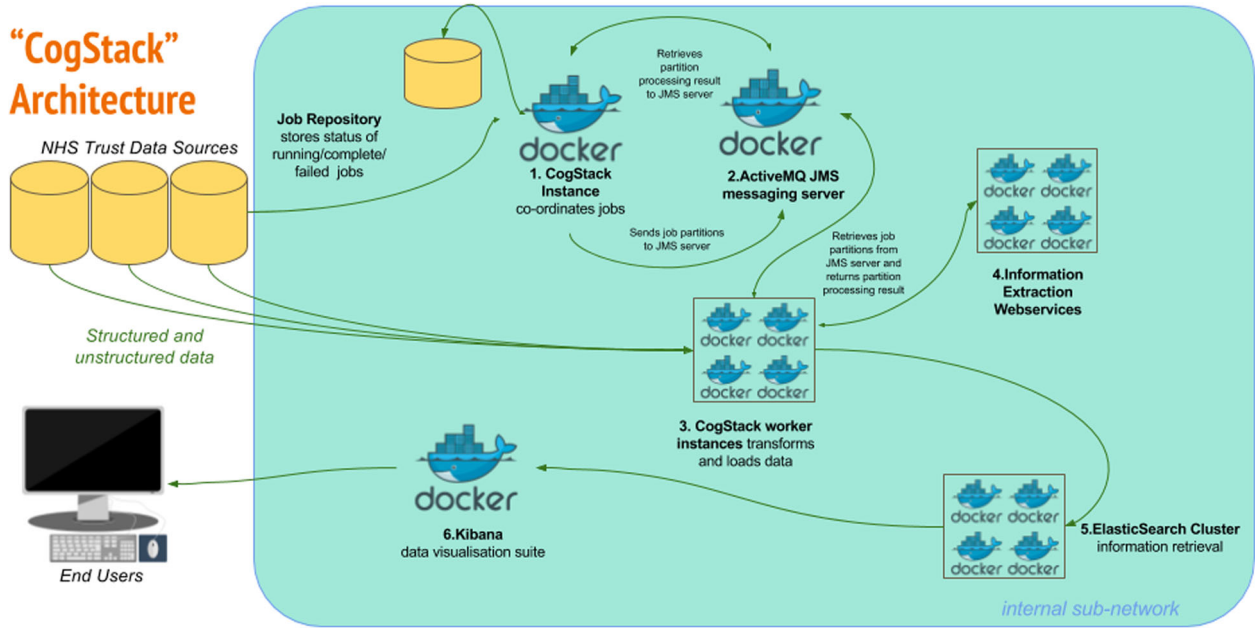


Fig. 1 CogStack Architecture and Dataflow All components can be deployed via the Docker containerisation software. **1. New job execution** Master instance of CogStack identifies new data in Trust Data Sources at intermittent intervals. **2. Partitioning** The job is partitioned into a user definable number of work units. **3a. Derive the freetext content** Extract plain and/or formatted text from common proprietary document binary formats (performing OCR where necessary), using the Tika Library to enable the downstream processing of high value unstructured data elements. **3b. Supplement the text content with meta-data** Filter and de-normalise a subset of the structured clinical data to provide a patient orientated, transparent representation of high value metadata concepts. For example, this might include calculated fields to represent patient age at document date, first part of postcode and ethnicity and lab results. **3c. De-identification** Transform the resulting text documents into de-identified text documents, by masking personal health identifiers via the use of the Cognition de-identification algorithms. This is necessary to address governance concerns associated with the secondary use of patient data. Identifiers in structured data can be excluded via SQL query, according to business requirements. **4. Information Extraction** Apply generic clinical IE pipelines to derive additional structured data from free text and supplement the quantity of available structured data at the point of query. **5. Indexing** Build a JSON object from the resulting structured and unstructured data, which can then be readily be indexed into an Elasticsearch cluster. **6. Visualisation** The Kibana suite provides a range of attractive options for viewing, aggregating and dash-boarding the loaded data

electronic free text content may be 'locked' inside proprietary file formats, or even lost during the conversion to an image format. The Apache Tika library [19] provides the capability to extract electronic text from a wide variety of file formats, and (in combination with the open source Tesseract Optical Character Recognition (OCR) tool [20]), recover images of text back into character electronic format. At the time of writing, Tika does not provide the capability to perform OCR on PDFs containing images. To this end, we enhance Tika with a custom PDF parser class, additionally making use of the ImageMagick tool in order to generate the required inputs for use with Tesseract.

Biomedical entity extraction, Bio-YODIE and Bio-LarK

Implementing an information retrieval system over clinical records represents a high return on investment by lowering the barrier to large scale data access in line with the 3R principle. However, the limitations of information retrieval are well recognised in terms of its ability to deal with ambiguity, different word senses, negation and other factors that are likely to produce an irrelevant or imprecise result. In order to provide a higher granularity of data at the point of search, it is necessary to implement information extraction (IE) techniques to enhance text elements with meta-data. To this end, the CogStack architecture offers two third party pipelines, with the capability to extend the system with additional pipelines via webservices.

First, Bio-YODIE (manuscript in preparation) is a clinical information extraction system designed for use with UK clinical records. It's development was necessitated in response to the widely recognised generalisability issues of English language clinical NLP systems, which have historically arisen in the United States [21, 22].

Bio-YODIE uses a configurable set of concepts from the Unified Medical Language System [23] Metathesaurus to provide natural language vocabularies of biomedical concepts, which it then attempts to disambiguate to UMLS concept unique identifiers. In this deployment of Bio-YODIE, we used all English

language concepts of the 2016AA release of the UMLS Metathesaurus.

Bio-YODIE has been evaluated against two corpora; the MIMIC II corpus [24, 25] and a new corpus created using patient records at the South London and Maudsley NHS Foundation Trust. In the latter, 201 documents have been triple-annotated by medical experts, achieving a three way interannotator agreement of 0.747. The corpus is confidential; however annotator guidelines are available for public review [26]. Bio-YODIE achieves an accuracy of 0.926 on the task of correctly linking to UMLS concepts on the SLAM corpus, 0.842 on the MIMIC 2013 test set and 0.827 on the MIMIC 2014 test set. A separate evaluation of NER performance (finding the right parts of the text, rather than as above, disambiguating them correctly given that the span has already been located) shows that Bio-YODIE achieves an F1 of 0.751 on perfect span matches (0.823 when concepts with any degree of overlap are also counted) on the SLAM corpus; however when only correct types are counted, this falls to 0.523 (0.564). NER performance was not evaluated on the MIMIC corpus because this corpus is not fully NER-annotated. In a comparative evaluation (forthcoming), Bio-YODIE and MetaMapLite offered similar advantages over the competitors considered in terms of accuracy, speed and stability; however, Bio-YODIE also offers the possibility to include prior probabilities from corpus data, resulting in a substantial improvement in disambiguation accuracy. For this reason, Bio-YODIE was chosen. Bio-YODIE is dual licensed under GNU Affero and commercial options.

Second, Bio-LarK encodes clinical text with Human Phenotype Ontology [27] concepts - the principle ontology for phenotyping patients in the 100K Genomics England Project. Negation detection for HPO terms is provided by the NegEx algorithm [28]. An evaluation of the system over a Pubmed corpus is described in [29]. Here, Bio-LarK achieved an F1 score of 0.95 over a test set of 1 933 instances, corresponding to 460 unique HPO concepts. Bio-LarK is available under an academic license.

The outputs of the NLP processes are captured as JSON objects and indexed using the 'nested' type of Elasticsearch. In doing so, it is possible to query unstructured data as though it were structured, although the accuracy will vary greatly depending on a multitude of factors.

Text de-identification performance

Different use cases for Trust data have different governance requirements. The requirements for the anonymisation and pseudonymisation has been the subject of national and international working groups [30–32]. The process of masking Protected Health Identifiers (PHIs) in clinical free text remains an active area of research from both a governance and NLP perspective. The Informatics for Integrating Biology and the Bedside (I2B2) organisation regularly organises open challenges for NLP researchers to examine the state of the art in text de-identification technology, by providing corpora of PHI annotated clinical text for international researchers to experiment with [33]. Such efforts have undoubtedly yielded significant advances in the field, to the extent that the performance of hybrid knowledge driven and machine learning methods equals that of human annotated documents in controlled test environments. Nevertheless, there remain outstanding tasks to ensure that such approaches are generalisable across different languages, dialects, specialities and hospital systems.

Due to strict data protection laws, it is generally not possible for researchers to access clinical text containing identifiable information. Therefore, validating the Cognition de-identification algorithms poses certain challenges. While certain domain corpora are available via activities such as I2B2 described above, these are not representative of UK clinical data. Therefore, we created a simulated dataset to explore the performance on registered company address entities harvested from public records. We devised a series of string mutator methods to attempt to recreate a variety of likely scenarios that would cause named entities to vary between two sources. These mutation methods were designed to represent real world events that might cause clinical document PHIs to not match those entered via an administrative process, and thus limit the effectiveness of exact string matching. We decided to focus on address named entities only, as these tend to offer the greatest scope for variation, compared with first/last name, telephone number and NHS number PHIs.

We explored four types of mutation method. First, keyboard typographic errors using prior probabilities of frequently mistyped keys, at a per character error rate of 3%, 10% and 20% (for example, '100 Meadow Street' to '100 Meagow Streat'. Second, substituting full address tokens to common abbreviations and vice versa (for example, 'Road' to 'Rd' and 'St' to 'Street') at a per token rate of 100% (i.e. any detected possible address substitutions were replaced). Third, an address token truncator, which removes tokens from the end of an address. The purpose of this is to replicate the observation that in some cases, full addresses (often supplemental address lines) are not recorded. For instance, '100 Meadow Street, Barkingford, Greater London, London' may be shortened to simply '100 Meadow Street'. We specified a token removal rate of 100%, with a minimum address length of three tokens. Finally, the most convoluted mutator we implemented was designed to mimic the effects of poor quality OCR. This mutator includes the effects of the character substitution mutator, with the additional possibility of inserting whitespace characters at random intervals within tokens. We tested this mutator with a per character substitution rate of 3%, 10% and 20%,

and a per character whitespace insertion rate of 3%, 10% and 20%

The mutated address strings were then wrapped in 'Lorem Ipsum' style generated text to simulate surrounding language. We generated 1000 test documents under a variety of degrees of PHI mutation and report precision and recall statistics for per token masking.

Scalability and database synchronisation

Scalability is achieved using the remote partitioning concept. Here, a unit of work is defined as a job (for example, to process 10 000 rows of new/updated data since the last job was executed). A master process partitions this job into a configurable number of smaller work units. These partitions and other job metadata are stored in a job repository and then sent as a message to a Java Messaging Service (JMS) server. These are then picked up by multiple worker processes operating on local or remote servers. Upon the arrival of a partition, each worker will begin to execute the work described within the message. Upon completion, the worker processes will inform the master process (again via JMS) about the status of the partition. If all partitions are successful, the job will be marked as complete, and a new job will start to process any new data generated by business activities during the processing of the previous job. Via this mechanism, a degree of 'near real-time' synchronisation with the source databases are achieved, although in practice it is constrained by available hardware, database configuration and network speed.

Elasticsearch and Kibana

Following the data transformation steps, the data is loaded into Elasticsearch, a popular open source search and analytics engine developed by Elastic.co. The non-transactional, NoSQL data model used by Elasticsearch enables the ingestion of large quantities of data at high speed, making it rapidly available for querying. Elasticsearch was chosen as it offers a number of advantages over traditional relational databases, predominantly concerning it's advanced capabilities to construct complex queries over structured and unstructured data simultaneously. In addition, the NoSQL data model it supports enables schema free loading of data (in the sense that there is no need to predefine the structure of data before it is loaded). This is particularly advantageous given the myriad of different database systems supported within a typical NHS Trust, as the technical debt incurred by connecting new data sources to the engine is greatly reduced. As an analytics engine, Elasticsearch allows common and complex aggregations to be performed at speed. Finally, Elasticsearch offers a Representational State Transfer (REST) web service, which cab be flexibly leveraged to allow external applications and services to retrieve data using the HTTP protocol.

For the end user experience, the open source Kibana data visualisation application (also by Elastic.co) is specifically designed to interact with Elasticsearch, and offers document visualisation, text highlighting and dashboarding capabilities. Via Kibana, non-technical users are able to search document text and structured metadata in much the same way as one would use an e-commerce website. A screen shot of the Kibana interface is provided in Fig. 2.

Security and information governance

Due to the sensitive nature of the clinical data, access is administered via a system manager in line with the information governance scenarios described above. Technical considerations are managed via commercial grade security provided by Elasticsearch plugins, offering Active Directory/LDAP/HTTP user authentication control, user access logging for audit, per index access restrictions with optional document/field level access restrictions and private certificate authority SSL encryption to protect in-flight data.

Results

Data model

As of December 2016, we have used the CogStack architecture to process approximately 300 million rows of clinical data from KCH databases. This data has been organised into identifiable and de-identified indexes for business intelligence and service improvement concerns, trial recruitment and tailored care use cases.

Each index is centred around a high value concept-

Observations Clinical notes taken during patient/doctor interactions (24 991 406 rows)

Basic observations Test results from pathology systems and short notes (248 028 823 rows)

Orders Prescribing information (66 838 164 rows)

Documents Binary documents generated by inter and intra Trust communication, comprising 8 736 295 rows. Of these, 4 505 750 (52%) resulted from MS Office, 3 479 583 (40%) were PDFs and 340 764 (4%)required OCR

Demographics of the acute patient population at King's College Hospital

A short demographic breakdown of patients across all years is given in Table 1. Top level ICD-10 groups, as assigned by clinical coders are presented in Table 2.

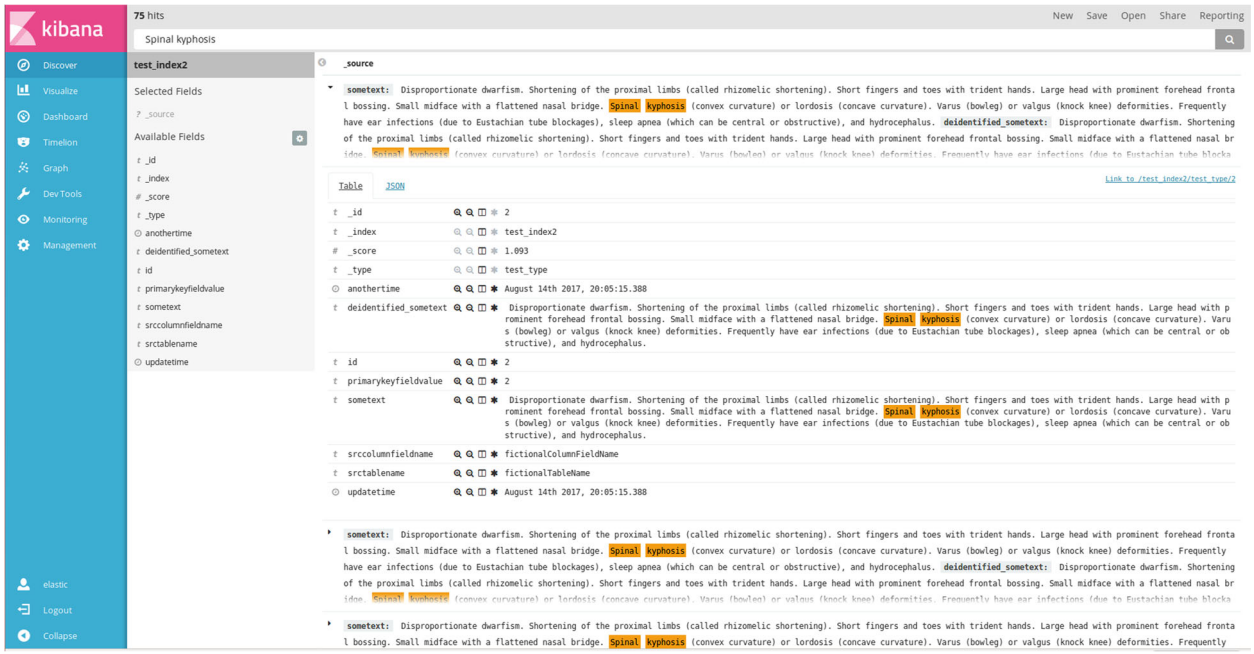


Fig. 2 Kibana interface loaded with pseudo-data

Table 1 Patient demographics, King's College Hospital 2004-2016

	Count	%
Age (years)		
≤ 20	435 796	14.80
21-40	811 865	27.57
41-60	876 467	29.77
61-80	490 153	16.65
≥ 80	326 453	11.09
Unknown	3 792	0.13
Gender		
Male	1 369 074	46.50
Female	1 571 717	53.38
Indeterminate	550	0.02
Unknown	3 185	0.11
Race (Self assigned		
Asian or Asian British	95 682	3.25
Black or Black British	326 618	11.09
Mixed	59 214	2.01
Not specified	1 506 703	51.17
Other	9 7277	3.30
White	859 032	29.17

Text de-identification validation

The results of our four methods to simulate PHI input errors for 1 000 addresses are given in Table 3. Because of the use of a random number generator to determine when string manipulations should occur, the total number of PHI tokens varies slightly between executions. For each test, approximately 8 500 pseudo-PHI address tokens were generated. For our character substitution mutator, precision ranged from 93.9% at a 3% substitution rate to 96.3% at a 20% substitution rate. Recall ranged from 95.5% at a 3% substitution rate to 82.0% at a 20% substitution rate. Performance over address aliasing achieved 94.4% precision and 94.8% recall. For token removal, precision was calculated at 96.6% and recall 92.1%. Performance on simulated OCR documents performed the least well, with precision at 98.2% and recall at 84.5% at a 3% character substitution rate and 3% white space insertion rate. At 20% charater substition rate and 20% white spcae insertion rate, precision was 92.3% and recall was 11.0%.

Discussion

Our CogStack software arose out of a requirement from the 100 000 Genome project (100KGP) to find a low cost solution for providing relevant clinical data to the programme amongst the large volumes of disparate data sources within KCH. In response to this challenge, we have produced an open source integrated document retrieval and information extraction, to solve a variety of typical issues associated with analytics within an

Table 2 ICD10 Code assignment by clinical coders at King's College Hospital

Group	Unique patient count
I Certain infectious and parasitic diseases	171 988
II Neoplasms	259 975
III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	72 939
IV Endocrine, nutritional and metabolic diseases	272 317
IX Diseases of the circulatory system	504 581
V Mental and behavioural disorders	706 990
VI Diseases of the nervous system	179 710
VII Diseases of the eye and adnexa	183 841
VIII Diseases of the ear and mastoid process	13 416
X Diseases of the respiratory system	242 282
XI Diseases of the digestive system	598 165
XII Diseases of the skin and subcutaneous tissue	131 227
XIII Diseases of the musculoskeletal system and connective tissue	343 803
XIV Diseases of the genitourinary system	212 198
XIX Injury, poisoning and certain other consequences of external causes	351 608
XV Pregnancy, childbirth and the puerperium	327 111
XVI Certain conditions originating in the perinatal period	78 541
XVII Congenital malformations, deformations and chromosomal abnormalities	104 242
XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	513 384
XX External causes of morbidity and mortality	650 984
XXI Factors influencing health status and contact with health services	1 520 107
XXII Codes for special purposes	385 417

NHS environment. We chose a range of components on the basis of ubiquity, robustness, commercially friendly licensing and price to offer a viable alternative to commercial solutions. Beyond the initial scope of the 100KGP, our CogStack architecture has enabled us to transform and ingest a large volume of clinical data in a fashion consistent with the requirements for data reuse in business intelligence, service improvement and research.

Case study: patient recruitment into the 100 000 Genomes England Project

The 100 000 Genomes project is the largest human sequencing project in the world. It is a UK initiative to sequence 100 000 genomes from individuals suffering from various cancers and rare diseases, with the intent of developing a genomic medicine capability for the NHS. This will create new diagnostic criteria for patients, and contribute to research for new treatments and cures. While the ambitions are high, the logistical and technical challenges of delivering such a capability within routine care are substantial. Two areas of particular difficulty have been identified at KCH.

The first challenge is to find and contact eligible patients for recruitment. Genomics Medicine Centres around the UK (such as KCH) are responsible for the recruitment and data collection of patients into the project, using the

Table 3 Performance of de-identification on simulated data

Mutator type	True positives	False positives	False negatives	Precision	Recall
Character substitution (3%)	8 191	538	391	93.9	95.5
Character substitution (10%)	7 740	447	826	94.6	90.4
Character substitution (20%)	6 969	271	1 537	96.3	82
Address Alias Substitution	8 171	486	455	94.4	94.8
Address Token Removal	2 761	99	237	96.6	92.1
OCR (3% char. sub. 3% white space	8 464	160	1555	98.2	84.5
OCR (10% char. sub. 10% white space	5 327	180	7282	96.8	42.3
OCR (20% char. sub. 20% white space	1 802	151	14719	92.3	11.0

various inclusion criteria specified by the co-ordinating body, Genomics England. One of the principal use cases for the CogStack architecture has been to offer the means to rapidly develop search criteria such that appropriate individuals are identified.

As noted by Moen et al. [12], quantitatively validating the quality of results produced by an information retrieval system is a complex task, as identifying the relevance of results is often highly context specific. However, subjective reports of users of the system suggest that project staff are able to work with clinical care teams to navigate large quantities of structured and unstructured data, to find information required validate putative cases for recruitment and approach patients in their normal course of care. For instance, the system allows researchers to quickly assess which patients have records that contain pertinent keywords and/or UMLS concepts, a process that would have previously required significant technical skill, direct knowledge of patient cases or manual data trawling.

The second challenge posed is to subsequently surface the deep phenotype data from recruited patients. A requirement for acceptance into the 100KGP is the completion of an extensive patient phenotypic data model by the recruiting Genomic Medical Centre. Such data may be held in disparate systems, complicating its extraction. Similarly to the recruitment challenge, collating data is substantially easier if held in a single source with extensive search functionality. In addition, the added value of IE approaches to resolve relevance challenges such as word sense disambiguation and negation offer further options for data retrieval. The technologies that make up the CogStack architecture enables members of the 100KGP team to rapidly scour individual patient records, regardless of size, and efficiently extract the required information.

Other use cases

While CogStack was built in response to the requirements of the 100k Genomics England project, its potential for a large number of other use cases was quickly realised. For instance, as previously described, clinical coding is the activity of hand curating clinical documents to identify the exercise of care activities, such as the prescription of drugs. Clinical coding is an important activity in acute care Trusts, as its efficiency affects the Trusts reimbursement from central government for care dispensed. The modern propensity to record and store large amounts of clinical and administrative data has created new challenges for clinical coders, owing to the increasingly unfavourable ratio of coding capacity to volume of data. Information retrieval and extraction technologies offer the potential for a substantial return on investment enabling clinical coders to navigate the data more efficiently. Such a capability is especially valuable in complex cases, where co-morbidity factors hidden amongst a mass of unstructured data can have a substantial impact in the accurate assessment of the cost of patient care.

In addition, one of the most useful tasks in an organisation with complex data flows is to be able to offer near real-time alerting capability. The commercial 'Alerting' plugin for Elasticsearch offers an easily configurable solution to send messages to a variety of endpoints, such as email addresses, REST webservices and enterprise communication software such as Slack and Hipchat. In a clinical setting, alerting clinical teams to events outside of their immediate jurisdiction may offer new opportunities for intervention. Within KCH, such capabilities are currently being explored in the following scenarios: 1) abnormal creatinine and CCP antibody levels to detect adverse reactions to methotrexate and pre-clinical rheumatoid arthritis respectively, to hasten communication between the Rheumatology and Pathology Departments 2) identification of previous evidence of adverse reactions such as rash in response to Sulfasalazine treatment (especially in emergency contexts) 3) monitoring for drug administration delays on wards 4) alerting of anti-coagulant team for patients being discharged on anti-coagulation therapy and 5) alerting of clinical intervention team if a high National Early Warning Score is detected. Presumably, such a list represents only a fraction of the scenarios that would benefit from the 3R principle. Pending further development and successful trials, a future goal will be to explore additional alerting scenarios.

Additional implementation issues/limitations

The secondary reuse of EHRs is complicated by several factors. Fundamentally, the clinotype and phenotype are related but different concepts in our semantics for heath datasets. The sufficiency and robustness of the clinical record is often called into question as a source of secondary research data [34–36]. For instance, our current deployment of CogStack at KCH does not have access to primary care data, and therefore cannot be said to offer complete patient profiles. Similarly, no effort has been made at this time to address the challenges of linking datasets across different secondary care organisations.

EHR data is predominantly used for front line recording and communication within care units. Missing data, or inconsistencies can be resolved if and when they become relevant to direct care by patient/care unit interaction. Such error correction routines are not possible in secondary use scenarios whereby corrective intervention is not feasible. The heterogeneous landscape of systems, data owners and APIs that are synonymous with IT infrastructure in large organisations are likely to compound the problem. Recording the same (or related) information in multiple systems, increases the likelihood of conflicts.

Governance, security and process issues require significant consideration in the development of standard operating procedures. It is likely that many Trusts have procedures in place to manage business intelligence, service improvement and research project with explicitly obtained consent. However, some of the most forward thinking opportunities for analytics require access to data at a scale where explicitly obtained consent is not feasible. Such activities likely require the use of external resources and expertise, as has been the doctrine behind the establishment of NHS/University collaborations in the form of National Institute for Health Research Biomedical Research Centres. Few Trusts have the facilities to offer enclave style research environments to external researchers, for example in the form of the aforementioned CRIS and SAIL security models. This creates a significant limitation in the potential for localised secondary EHR use outside of such institutions, and discussions to address such issues continue to take place at the national level. Progress in this area is likely to take the form of substantial patient engagement activities to ensure the retention of public trust, and the development of pioneering models of consent such as Consent for Contact [37].

One particular factor of concern when managing unstructured data is the quality of OCR performance. Although only 4% of binary documents required OCR at KCH, our subjective assessment of the Tesseract library suggests OCR performance varies greatly in line with the quality of input. Good performance was observed when OCR was attempted on clean, printed black and white documents that were carefully aligned to scanner boarders. Deviations from these factors resulted in a rapid decline in OCR performance.

Regarding Information Extraction approaches, our efforts here offer Bio-LarK and Bio-YODIE 'out-of-the-box' as a means to demonstrate compatibility with the CogStack concept. However, the necessity for domain adaption to new corpora of clinical text is well established [38, 39]. Future work will look at the information extraction performance and ease of domain adaptation of these technologies to the KCH corpora.

The de-identification algorithms we make use of are deterministic string matching method based upon the same principles described in [17]. Although we were unable to validate the performance on real clinical data at this time, we would expect recall metric to be approximately the same.

Because of our access limitations to identifiable clinical data, we are hesitant to make broad comparisons with other methods in this area. We would have liked to compare performance across a range of algorithms, such as those proposed in the I2B2 2014 task for text de-identification. However, it should be noted that the majority of these algorithms are not available in the public domain. In addition, we note that such algorithms are designed for US style identifiers rather than UK ones, therefore requiring some form of domain adaptation for appropriate use. Regardless, our experiences of automated de-identification techniques suggest that appropriate ethical use should involve extensive internal validation on a per-dataset basis, before such data is deemed suitably transformed for further use cases.

Our testing of the approach in a simulated environment suggests reasonable performance of the de-identification algorithms to many forms of string perturbation, with the most noticeable drops in performance occurring with our 'poor OCR' simulations. It should be noted that at the higher grades of OCR error, documents became increasingly illegible, suggesting that PHIs may not be interpretable to human observers.

One particular dependency of the de-identification algorithm is that it requires PHIs to exist as structured or semi-structured fields in a database, which may make it unsuitable for some types of EHR data. Many other forms of PHI masker do not have this requirement [40]. However, due to the nature of its workings, it can synergistically be combined with other de-identification approaches.

Regarding resource allocation during the progress of the project, the most significant deployment cost arose from the need for the implementation team to understand the complex landscape of modern and legacy systems in place inside the Trust. For instance, these commonly took the form of certain services being unavailable at certain times, or restrictions on the load that could be placed on certain services to prevent interference with the day-to-day running of front line services. In such cases, it was necessary to retain flexibility with regard to requirements, in keeping with common agile management paradigms.

Conclusions

Our CogStack software arose out of a requirement to build an integrated document retrieval and information extraction system for a large UK NHS Trust. Our experiences have led us to identify a variety of typical issues associated with the development of local analytics environments within the NHS, broadly encapsulated as what we define as the 3Rs of right data, right place and right time. We have released our software components under permissive licensing arrangements in the hope that other NHS Trusts might benefit from our findings.

Availability and requirements

Project name: CogStack

Project home page: The code, documentation, string mutator classes and example configurations for CogStack

as describe in this article are available at https://github.com/RichJackson/cogstack/.

The latest version of CogStack can be found at https://github.com/cogstack/cogstack/

Operating system(s): JVM based - The codebase should work on Windows and Linux systems, although Linux systems are recommended for docker style deployment

Programming Language: Java, Groovy, Spring Batch Framework

Other requirements: Java 1.8 or higher

License: Apache 2.0

Any restrictions to use by non-academics: Please check with Angus Roberts (angus.roberts@sheffield.ac.uk) and Tudor Groza (t.groza@garvan.org.au) before using the Bio-YODIE and Bio-LarK components respectively

Abbreviations

100KGP: 100 000 genomes project; API: Application programming interface; CCP: Cyclic citrullinated Peptide; CRIS: Clinical records interactive search; EHR: Electronic health record; HL7: Health level 7; HPO: Human phenotype ontology; HTTP: Hypertext transfer protocol; IE: Information extraction; I2B2: The informatics for integrating biology and the bedside; JMS: Java messaging service; JSON: JavaScript object notation; KCH: King's college London; LDAP: Lightweight directory access protocol; NER: Named entity recognition; NHS: National health service; NLP: Natural language processing; OCR: Optical character recognition; OLAP: Online analytical processing; PHI: Protected health identifiers; REST: Representational state transfer; SAIL: Secure anonymised data link; SLAM: South London and Maudsley; SQL: Structured query language; SSL: Secure socket layer

Acknowledgements

We would like to thank Dr Will Bernal, Caldicott Guardian for KCH for his advice on governance and ethical matters.

Funding

This paper represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, the University College London Hospitals Biomedical Research Centre, by awards establishing the Farr Institute of Health Informatics Research at UCL Partners, from the Medical Research Council, Arthritis Research UK, British Heart Foundation, Cancer Research UK, Chief Scientist Office, Economic and Social Research Council, Engineering and Physical Sciences Research Council, National Institute for Health Research, National Institute for Social Care and Health Research, and Wellcome Trust (grant MR/K006584/1).
Additional funding came from the following awards. NHS England Enablement funding, the UK Infrastructure for Large-scale Clinical Genomics Research $MC_PC_1$4089 and the European Union's Horizon 2020 research and innovation programme under grant agreement No 644753 (KConnect).
The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.
The funding bodies had no role in the design of the study and collection, analysis, the interpretation of data nor in writing the manuscript.

Availability of data and materials

Instructions on how to reproduce the simulated data and results described here are included in the documentation of the CogStack repository, available at https://github.com/cogstack/cogstack/. For information governance reasons, no clinical record data can be provided with this research.

Authors' contributions

RJ led the architecture for CogStack, developed the batch process, data simulation algorithms, and the analysis of the data. RJ and IK developed the business requirements for Cognition. IK led the design and development of Cognition, including the de-identification algorithms used here. KL, AF and AA contributed to concept design and technical support. DL, DN and CS were responsible for the acquisition of the data and business support for KCH. TG was responsible for the Bio-LarK concept. AR, GG, XS and HW were responsible for the Bio-YODIE component. RS and RD contributed to the analysis and interpretation of the results. All authors were involved in manuscript preparation, approved the final version and agree to be accountable for the work.

Ethics approval and consent to participate

The creation of the CogStack software was an internal service development project for King's College Hospital NHS Foundation Trust, and thus did not require ethical approval. As no patient identifiable data was required for the development of the software, no approval was sought from the Health Research Authority according to Confidentiality Advisory Group guidelines (http://www.hra.nhs.uk/resources/confidentiality-advisory-group/determining-need-cag-application/). The validation of the Bio-YODIE software made use of the CRIS dataset, which is approved as an anonymised data resource for secondary analysis by Oxfordshire Research Ethics Committee C (08/H0606/71) and governance is provided for all projects and dissemination through a patient-led oversight committee.

Consent for publication

Not applicable: No individual persons data is presented in this manuscript.

Competing interests

RJ and RS have received research funding from Roche, Pfizer, J&J and Lundbeck.

Publisher's Note

Author details

[1]Institute of Psychiatry, Psychology and Neuroscience, King's College London, 16 De Crespigne Park, SE5 8AF London, UK. [2]South London and Maudsley NHS Foundation Trust, Denmark Hill, SE5 8AZ London, UK. [3]King's College Hospital, Denmark Hill, SE5 9RS London, UK. [4]University of Sheffield, Western Bank, S10 2TN Sheffield, UK. [5]Farr Institute of Health Informatics Research, UCL Institute of Health Informatics, University College London, WC1E 6BT London, UK. [6]Garvan Institute of Medical Research, NSW 2010 Sydney, Australia. [7]InterDigital Communications, 64 Great Eastern Street, 1st Floor, EC2A 3QR London, UK. [8]Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, EH16 4UX Edinburgh, UK.

Received: 20 March 2017 Accepted: 1 June 2018
Published online: 25 June 2018

References

1. Simborg DW. An emerging standard for health communications: The HL7 standard. Healthc Comput Commun. 1987;4(10):58–60.
2. Klein GO. Standardization of health informatics–results and challenges. Methods Inf Med. 2002;41(4):261–70.
3. Barnes M. Lessons learned from the implementation of clinical messaging systems. AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium. Montgomery: The American Medical Informatics Institution; 2007, pp. 36–40.
4. Worden R, Scott P. Simplifying HL7 Version 3 messages. Stud Health Technol Inform. 2011;169:709–13.
5. Antolík J. Automatic annotation of medical records. Stud Health Technol Inform. 2005;116:817–22. Cited by 0003.
6. Albright D, Lanfranchi A, Fredriksen A, Styler WF, Warner C, Hwang JD, Choi JD, Dligach D, Nielsen RD, Martin J, Ward W, Palmer M, Savova GK. Towards comprehensive syntactic and semantic annotations of the clinical narrative. J Am Med Inform Assoc. 2013;20(5):922–30.
7. Barrett N, Weber-Jahnke JH. Applying natural language processing toolkits to electronic health records - an experience report. Stud Health Technol Inform. 2009;143:441–6.
8. Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: A description based on the theories of Zellig Harris. J Biomed Inform. 2002;35(4):222–35.
9. Perera G, Broadbent M, Callard F, Chang C-K, Downs J, Dutta R, Fernandes A, Hayes RD, Henderson M, Jackson R, Jewell A, Kadra G, Little R, Pritchard M, Shetty H, Tulloch A, Stewart R. Cohort profile of the

South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: Current status and recent enhancement of an Electronic Mental Health Record-derived data resource. BMJ Open. 2016;6(3):008721.
10. Jones KH, Ford DV, Jones C, Dsilva R, Thompson S, Brooks CJ, Heaven ML, Thayer DS, McNerney CL, Lyons RA. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: A privacy-protecting remote access system for health-related research and evaluation. J Biomed Inform. 2014;50:196–204.
11. The 100,000 Genomes Project Protocol v3, Genomics England. 2017. https://doi.org/10.6084/m9.figshare.4530893.v2. (from https://www.genomicsengland.co.uk/about-gecip/publications/).
12. Moen H, Ginter F, Marsi E, Peltonen L-M, Salakoski T, Salanterä S. Care episode retrieval: Distributional semantic models for information retrieval in the clinical domain. BMC Med Inform Dec Making. 2015;15(S2).
13. McEwan R, Melton GB, Knoll BC, Wang Y, Hultman G, Dale JL, Meyer T, Pakhomov SV. NLP-PIER: A Scalable Natural Language Processing, Indexing, and Searching Architecture for Clinical Notes. AMIA Jt Summits Transl Sci Proc. 2016;2016:150–9.
14. Stewart R, Soremekun M, Perera G, Broadbent M, Callard F, Denis M, Hotopf M, Thornicroft G, Lovestone S. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: Development and descriptive data. BMC Psychiatry. 2009;9(1):51.
15. Kartoglu IE. Cognition: DB binary-to-text converter and pseudonymiser for clinical research. 2015. https://github.com/KHP-Informatics/Cognition-DNC.
16. Jackson R, Kartoglu I. A Open Pipeline for Masking Patient Identifiers in Electronic Health Records, The Farr Institute International Conference 2015. 2015.
17. Fernandes AC, Cloete D, Broadbent MT, Hayes RD, Chang C-K, Jackson RG, Roberts A, Tsang J, Soncul M, Liebscher J, Stewart R, Callard F. Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. BMC Med Inf Decis Making. 2013;13(1):71.
18. D'Avolio LW, Nguyen TM, Goryachev S, Fiore LD. Automated concept-level information extraction to reduce the need for custom software and rules development. J Am Med Inform Assoc. 2011;18(5):607–13.
19. Mattmann C, Zitting J. Tika in Action. Greenwich, CT, USA: Manning Publications Co.; 2011.
20. Smith R. An Overview of the Tesseract OCR Engine. In: Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02. ICDAR '07. Washington, DC, USA: IEEE Computer Society; 2007. p. 629–33.
21. Wu S, Miller T, Masanz J, Coarr M, Halgrim S, Carrell D, Clark C. Negation's Not Solved: Generalizability Versus Optimizability in Clinical Natural Language Processing. PLoS ONE. 2014;9(11):112774.
22. Friedman C, Hripcsak G. Evaluating natural language processors in the clinical domain. Development. 1998;22:24.
23. Bodenreider O. The unified medical language system (UMLS): Integrating biomedical terminology. Nucleic Acids Res. 2004;32(suppl 1):267–70.
24. Neamatullah I, Douglass MM, Lehman L-wH, Reisner A, Villarroel M, Long WJ, Szolovits P, Moody GB, Mark RG, Clifford GD. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak. 2008;8(1).
25. Johnson AEW, Pollard TJ, Shen L, Lehman L-wH, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035.
26. University of Sheffield. KConnect UMLS Annotation Task. 2016. http://www.dcs.shef.ac.uk/~genevieve/kconnect/annotation-manual.pdf.
27. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GCM, Brown DL, Brudno M, Campbell J, FitzPatrick DR, Eppig JT, Jackson AP, Freson K, Girdea M, Helbig I, Hurst JA, Jähn J, Jackson LG, Kelly AM, Ledbetter DH, Mansour S, Martin CL, Moss C, Mumford A, Ouwehand WH, Park S-M, Riggs ER, Scott RH, Sisodiya S, Vooren SV, Wapner RJ, Wilkie AOM, Wright CF, Vulto-van Silfhout AT, de Leeuw N, de Vries BBA, Washingthon NL, Smith CL, Westerfield M, Schofield P, Ruef BJ, Gkoutos GV, Haendel M, Smedley D, Lewis SE, Robinson PN. The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. Nucleic Acids Res. 2014;42(D1):966–74.
28. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. J Biomed Inform. 2001;34(5):301–10.
29. Groza T, Kohler S, Doelken S, Collier N, Oellrich A, Smedley D, Couto FM, Baynam G, Zankl A, Robinson PN. Automatic concept recognition using the Human Phenotype Ontology reference and test suite corpora. Database. 2015;2015(0):005.
30. Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. J Biomed Inform. 2015;58:20–29.
31. Aamot H, Kohl CD, Richter D, Knaup-Gregori P. Pseudonymization of patient identifiers for translational research. BMC Med Inform Dec Making. 2013;13(1).
32. Crown. Anonymisation: Managing Data Protection Risk Code of Practice, Infomation Commisioners Office. 2012. https://ico.org.uk/media/1061/anonymisation-code.pdf.
33. Malin BA, Emam KE, O'Keefe CM. Biomedical data privacy: Problems, perspectives, and recent advances. J Am Med Inform Assoc. 2013;20(1):2–6.
34. Munk-Jørgensen P, Okkels N, Golberg D, Ruggeri M, Thornicroft G. Fifty years' development and future perspectives of psychiatric register research. Acta Psychiatr Scand. 2014;130(2):87–98.
35. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: Data quality issues and informatics opportunities. AMIA Summits Transl Sci Proc. 2010;2010:1–5.
36. Mikkelsen G, Aasly J. Consequences of impaired data quality on information retrieval in electronic patient records. Int J Med Inform. 2005;74(5):387–94.
37. Callard F, Broadbent M, Denis M, Hotopf M, Soncul M, Wykes T, Lovestone S, Stewart R. Developing a new model for patient recruitment in mental health services: A cohort study using Electronic Health Records. BMJ Open. 2014;4(12):005654.
38. Ferraro JP, Daume H, DuVall SL, Chapman WW, Harkema H, Haug PJ. Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. J Am Med Inform Assoc. 2013;20(5):931–9.
39. Zhang Y, Tang B, Jiang M, Wang J, Xu H. Domain adaptation for semantic role labeling of clinical text. J Am Med Inform Assoc. 2015;22(5):967–79.
40. Stubbs A, Kotfila C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. J Biomed Inform. 2015;58:11–9.

6.3 Conclusion

The CogStack is an example of the possibilities that are created via successful NHS/academic collaboration. In keeping with the motivation of this thesis, the CogStack project has been designed for widespread adoption throughout the NHS. Although the nature of the project is such that only a suitably equipped NHS I.T. department will be able to make use of it, it is designed to lower the barrier for successful exploitation of clinical text resources without the need for users to have an extensive NLP background. As evidence of its success in this regard, CogStack has since gained traction within other NHS organisations, with funding secured for deployments at Norfolk and Norwich University Hospitals NHS Foundation Trust, University College London Hospital NHS Foundation Trust, and (at time of writing) further discussions underway for a deployment at Guy's and St Thomas' NHS Foundation Trust. Its impact has also been acknowledged in the Annual Report of the Chief Medical Officer 2016 [246]. The range and scope of projects utilising the CogStack at King's College Hospital continues to grow, with the hope that it will eventually form part of a new research enclave at King's College Hospital Foundation Trust.

Chapter 7

Discussion

The goals of this thesis included exploring ways to bridge the gap between some aspects of NLP in theory and real world clinical applications. Early work on an exploration of IE techniques to extract a selection of the negative symptoms of SMI suggested that SVMs were able to produce good classification performance in this task, with only relatively small amounts of training data. This set the scene for scaling this approach to attempt a far larger number of symptoms, via the CRIS-CODE project. First, however, it was necessary to develop the TextHunter tool to streamline the required processes. 50 separate symptoms were tackled in CRIS-CODE, necessitating over 40 000 annotations, but in return yielding the symptom profiles for 7 962 SMI patients. The outputs of CRIS-CODE and TextHunter have already formed the bases of several follow on efforts, including an examination of the clinical outcomes of mood instability [11], the relationship between cannabis use and outcomes in first episode psychosis [10], and the relationship between SMI symptom profile and length of inpatient stay (manuscript in preparation).

Although the TextHunter software is far from perfect, its popularity stems from two key features, tackling the business requirement of lowering the technical barrier for effective IE pipeline creation. First, the implementation of a simple, efficient annotation interface enables users without a core interest in NLP to forgo technical training in packages such as GATE or NLTK, in order to harness the value of IE. The second key feature of TextHunter was the implementation of a hybrid rules and ML pipeline for concept extraction, which was generalisable enough such that it could be applied with frequent success across many concepts. Although our results suggest a generalisable IE methodology for any concept remains elusive, we were able to achieve acceptable classification performance for most of the symptoms tackled in CRIS-CODE. The main methodological finding from this work suggested that rules-based approaches generalise well to a degree (such as with the ConTEXT algorithm), after which further development yields a diminishing return on

investment. Here, the value of ML takes over, and tends to increase classification performance without requiring language engineering domain knowledge (a significant bonus in terms of skill availability in the labour force).

Nevertheless, in order to execute the CRIS-CODE project, we were required to make some assumptions about the clinical language of SMI symptomatology. These assumptions predominantly included the words and phrases in use around SMI symptom concepts (backed by their subjective observation in this CRIS system). Following CRIS-CODE, I chose to challenge this assumption, by attempting to discover the language that clinicians use to describe symptomatology 'in their own words'. By using the CBOW language model and the K-means++ clustering algorithm, I was able to find a range of new concepts and phrases in use that expanded upon traditional depictions, models and groupings of SMI symptoms. The exploration of their clinical significance will require substantial further work. However, the discovery of their existence within clinical databases suggests further interrogation of the raw data in large clinical corpora may prove fruitful.

The role of high quality software development is becoming increasingly important in the academic domain [247], yet the public discourse around such issues remains marginal. Having addressed the fundamentals of IE and knowledge discovery around SMI symptomatology in CRIS, my research shifted focus to my experiences of operationalising NLP processes in production. The problem of scalability was partly addressed in the TextHunter tool, although some design decisions led to subsequent limitations around portability and distributed computing. The CogStack project was therefore designed to address such problems of scalability in a general way, such that NLP solutions (including the TextHunter SMI models) could be deployed within resource constrained NHS environments. Documenting the development and maintenance of the CogStack offered the opportunity to describe important aspects of process that underpin many NLP research projects, yet are often neglected in methodological discussions. CogStack's success at King's College Hospital, with much assistance from the project sponsor, Clive Stringer, has gathered momentum in a new way of thinking about business intelligence within the Trust, and continues to gather additional use cases and interest (https://www.youtube.com/watch?v=XU9ls9J9AmM). The development of CogStack remains active.

7.1 Conclusion - The Role of NLP in the Distant Spectacle of an AI Doctor

In chapter 3, I described the motivation for this thesis as an exploration of clinical NLP as a solution to certain business problems in the clinical environment. I conclude that general

purpose clinical NLP systems are elusive with respect to historic data. Nevertheless, if they were to exist, one prominent use case for them would be as a component of a fully automated clinical decision making system - a clinical artificial intelligence. To conclude, I describe a personal opinion, looking to the wider role of NLP in ushering in the so called 'fourth industrial revolution, the age of Artificial Intelligence', with respect to healthcare.

There is an ongoing lively debate on forums and blogs about the role of Artificial Intelligence (AI) in the future of society. Amongst the more levelled predictions, discussions range from fears about unprecedented economic threats to the role of the worker as automation transforms the nature long-standing job classes, to ushering in a new age of economic prosperity where complex industries such as healthcare are democratised and made cheap and available for all. Whatever the future holds, NLP will almost certainly form a core component, as the rate of transfer of human knowledge and state to a machine representation becomes the limiting factor to further progress.

The domain of NLP at large continues to evolve at a rapid pace. In 2013, around the time the objectives for this thesis were being agreed, Tomas Mikolov and his collegues released two papers [248,249], describing the efficient training of neural networks in language modelling. Similarly in 2016, scientists at Google announced their creation of the most accurate syntatic parser ever devised, again using neural networks [250]. That the use of neural networks in NLP would enjoy such popularity in recent years might have been predicted, given the ground breaking success stories in computer vision, self driving cars, voice-to-text and similar technologies that may come to define the next era of the digital economy. Looking forward, it seems likely that neural networks and their interpretation will continue to play a significant role the research agenda in NLP for the imminent future, whereas the development of the user experience and subsequent commoditisation of such models will define the objectives of the business agenda.

Such work might be said to represent the cutting edge of NLP (or at least the current fashion). However, perhaps progress towards 'AI doctors' is not as forthcoming as the headlines would have us believe. Mikolov et als work is considered something of a flashpoint in the field, igniting a surge of interest in the concept of word embeddings. However, in terms of solving the problem of conveying the semantic meaning of words to a machine, the CBOW model is simply another way of representing relationships between words in vector space, a concept which had been around for many years. As for Google's parser, the accuracy improvement over the previous best effort is only in the order of a few percent. Yet there remains a weight of expectation that NLP will enable machines to interact as fully integrated 'AIs' in human society within a generation. In the popular science book 'The Rise of the Robots' [251], futurist Martin Ford describes at length the extent to

which narrative can be generated by machines, particularly noting the Quill technology by Narrative Science. Amongst its achievements, the Quill engine has authored a number of short articles for Forbes magazine and is listed as a 'Contributing Author'. Amongst the claims by the company's CTO were that by 2026, 90% of all news would be written by computer programs[1]. To counter the hyperbole and make reasonably informed predictions about the timescale of future progress, it is important to contextualise such advances in relation to the long list of technical milestones that lie ahead.

Let's consider the nature of human communication and the implications for human/-machine interface. Humans are 'messy' communicators. Discourse often need to be repeated, described other ways, demonstrated with examples and analogies before meaning is (hopefully) transferred. Intuitively, few people are able to self-learn complex subjects efficiently, necessitating the existence of an education system in most countries. Natural language, especially written language is only an imperfect vehicle for the abstract concept of human communication. This seems to be implicitly understood by humans, and, despite the modern tendency of accumulating vast amounts of textual data, most people prefer to communicate important topics in person, where the array of social cues such as tone of voice, facial expressions and other forms of body language afford a greater degree of expressiveness.

Further, 'meaning' associated with language changes as a function of time. Natural language might only be accurate in the 'present' - the interpersonal conversation between a set of individuals. In future time, the analysis of the transcript will undoubtedly change it's interpretation. In 2011, the Oxford dictionary acknowledged a particularly vivid example wherein 'literally' literally no longer just meant 'literally':

"*... ...In recent years an extended use of literally (and also literal) has become very common, where literally (or literal) is used deliberately in non-literal contexts, for added effect, as in they bought the car and literally ran it into the ground.* "

Fundamentally, interpersonal communication is almost always underwritten by the possibility of engaging in energetically expensive but less error prone multiscale methods (elaboration, demonstration, body language, facial expressions etc.), if meaning and/or intent is not transferred and acknowledged via more efficient routines. A large part of the practice of medicine at the point of care is an act of human communication. If a machine is unable to convey the full range of multiscale human communication, how will an AI doctor interact with its patients? For clinical NLP and in turn clinical AI to realise its

[1]Narrative Machines seems to have made its last contribution in October 2015

promises, natural language communication between patient and machine needs to be as fluid as communication between patient and doctor. For now, let's regard such a milestone as being in the distant future.

A more likely scenario that Ford speculates upon regards human intermediaries into the computer interface:

"*... Once machines can demonstrate that they can offer accurate diagnoses and effective treatment plans, perhaps it will not be necessary for a physician to directly oversee every encounter with every patient [...] there may eventually be an opportunity to create a new class of medical professionals: persons educated with perhaps a four-year undergraduate or master's degree, and who are trained primarily to interact with and examine patients - and then convey that information into a standardized diagnostic and treatment system.* "

Such a position seems feasible - indeed, clinical decision support systems for narrowly defined problems have been garnering attention for several years. However, the development of an AI doctor, taken to mean a machine that performs at least as well as a high calibre human doctor (most likely using some form of neural network) presents fundamental problems:

For brevity, let's consider a single aspect: the underlying ontology such a machine would require to represent the medical knowledge from which it would source the logic behind its decisions. To produce an ontology equal to the training and experience acquired by a single domain consultant over a career, is as much of a challenge in interpersonal communication as it is in ontology engineering. NLP is often proposed to automate the process to a degree, by mapping the semantics of human knowledge contained in literature resources into an appropriate ontological structure. However, once again, resolving fundamental differences in how humans internally represent 'meaning' and our machine equivalents create problems of the utmost difficulty. As Bimson and Hull note in a chapter aptly titled 'Unnatural Language Processing' [252]:

"*... The semantic expressiveness of ontologies simply is not sufficient to represent the semantic complexity of [natural language], at least not without building significant "representational scaffolding" to support it, leading to severe language-to-ontology mapping and modelling challenges. These challenges lead, in turn, to problems in extracting knowledge from text sources and representing it as ontology constructs. To borrow an analogy from the film industry, editing [natural language] semantics enough to fit into standard ontology structures requires us to leave a significant amount of valuable knowledge on the editing*

room floor. Understanding the semantic tradeoffs that must be made is critical for customers, information architects, and users, because meaning will be lost in the process of transforming [natural language] semantics into ontology semantics, meaning that is often important to stakeholders. "

Simply put, our best methods of modelling reality *in silico* lag far behind the unknown mechanisms at work in the human brain. Yet it is the construction of these very ontologies that present the most immediate roadblock to progressing the concept of anything approaching an AI doctor, and the resolution of such issues seems likely to be the focal task for at least the next couple of decades.

In conclusion, our current state of the art approaches to AI are many orders of magnitude less complex than the brains of even simple organisms. Nevertheless, the pursuit of AI in healthcare will likely sequester a large portion of private and public health informatics funding in years to come. Far from democratising medicine, failing to recognise the scope of the challenge may result in the opposite - the economic attractiveness of AI administered healthcare may divert resources from attempts to bolster higher quality, human mediated solutions. Our guiding principle towards an AI doctor should follow the ancient mantra: *Primum non nocere* - "First, do no harm".

Bibliography

[1] G. Perera, M. Broadbent, F. Callard, C.-K. Chang, J. Downs, R. Dutta, A. Fernandes, R. D. Hayes, M. Henderson, R. Jackson, A. Jewell, G. Kadra, R. Little, M. Pritchard, H. Shetty, A. Tulloch, and R. Stewart, "Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: Current status and recent enhancement of an Electronic Mental Health Record-derived data resource," *BMJ Open*, vol. 6, p. e008721, Mar. 2016.

[2] N. England, "Five Year Forward View." `https://www.england.nhs.uk/wp-content/uploads/2014/10/5yfv-web.pdf`. Accessed: 2016-08-30.

[3] R. G. Jackson MSc, M. Ball, R. Patel, R. D. Hayes, R. J. B. Dobson, and R. Stewart, "TextHunter - A User Friendly Tool for Extracting Generic Concepts from Free Text in Clinical Research," *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, vol. 2014, pp. 729–738, 2014.

[4] R. G. Jackson, R. Patel, N. Jayatilleke, A. Kolliakou, M. Ball, G. Gorrell, A. Roberts, R. J. Dobson, and R. Stewart, "Natural language processing to extract symptoms of severe mental illness from clinical text: The Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project," *BMJ Open*, vol. 7, p. e012012, Jan. 2017.

[5] R. Jackson, R. Patel, S. Velupillai, G. Gkotsis, D. Hoyle, and R. Stewart, "Knowledge discovery for Deep Phenotyping serious mental illness from Electronic Mental Health records," *F1000Research*, vol. 7, p. 210, May 2018.

[6] R. Jackson, I. Kartoglu, C. Stringer, G. Gorrell, A. Roberts, X. Song, H. Wu, A. Agrawal, K. Lui, T. Groza, D. Lewsley, D. Northwood, A. Folarin, R. Stewart, and R. Dobson, "CogStack - experiences of deploying integrated information retrieval and extraction services in a large National Health Service Foundation Trust hospital," *BMC Medical Informatics and Decision Making*, vol. 18, Dec. 2018.

[7] J. Downs, M. Hotopf, T. Ford, E. Simonoff, R. G. Jackson, H. Shetty, R. Stewart, and R. D. Hayes, "Clinical predictors of antipsychotic use in children and adolescents with autism spectrum disorders: A historical open cohort study using electronic health records," *European Child & Adolescent Psychiatry*, vol. 25, pp. 649–658, June 2016.

[8] A. Kolliakou, M. Ball, L. Derczynski, D. Chandran, G. Gkotsis, P. Deluca, R. Jackson, H. Shetty, and R. Stewart, "Novel psychoactive substances: An investigation of temporal trends in social media and electronic health records," *European Psychiatry: The Journal of the Association of European Psychiatrists*, vol. 38, pp. 15–21, Oct. 2016.

[9] R. Patel, H. Shetty, R. Jackson, M. Broadbent, R. Stewart, J. Boydell, P. McGuire, and M. Taylor, "Delays to diagnosis and treatment in patients presenting to mental health services with bipolar disorder," *European Psychiatry*, vol. 33, p. S75, Mar. 2016.

[10] R. Patel, R. Wilson, R. Jackson, M. Ball, H. Shetty, M. Broadbent, R. Stewart, P. McGuire, and S. Bhattacharyya, "Association of cannabis use with hospital admission and antipsychotic treatment failure in first episode psychosis: An observational study," *BMJ open*, vol. 6, no. 3, p. e009888, 2016.

[11] R. Patel, T. Lloyd, R. Jackson, M. Ball, H. Shetty, M. Broadbent, J. R. Geddes, R. Stewart, P. McGuire, and M. Taylor, "Mood instability is a common feature of mental health disorders and is associated with poor clinical outcomes," *BMJ Open*, vol. 5, pp. e007504–e007504, May 2015.

[12] R. D. Hayes, J. Downs, C.-K. Chang, R. G. Jackson, H. Shetty, M. Broadbent, M. Hotopf, and R. Stewart, "The Effect of Clozapine on Premature Mortality: An Assessment of Clinical Monitoring and Other Potential Confounders," *Schizophrenia Bulletin*, vol. 41, pp. 644–655, May 2015.

[13] R. Patel, N. Jayatilleke, M. Broadbent, C.-K. Chang, N. Foskett, G. Gorrell, R. D. Hayes, R. Jackson, C. Johnston, H. Shetty, A. Roberts, P. McGuire, and R. Stewart, "Negative symptoms in schizophrenia: A study in a large clinical sample of patients using a novel automated method," *BMJ Open*, vol. 5, p. e007619, Sept. 2015.

[14] E. Iqbal, R. Mallah, R. G. Jackson, M. Ball, Z. M. Ibrahim, M. Broadbent, O. Dzahini, R. Stewart, C. Johnston, and R. J. B. Dobson, "Identification of Adverse Drug

Events from Free Text Electronic Patient Records and Information in a Large Mental Health Case Register," *PLOS ONE*, vol. 10, p. e0134208, Aug. 2015.

[15] G. Kadra, R. Stewart, H. Shetty, R. G. Jackson, M. A. Greenwood, A. Roberts, C.-K. Chang, J. H. MacCabe, and R. D. Hayes, "Extracting antipsychotic polypharmacy data from electronic health records: Developing and evaluating a novel process," *BMC Psychiatry*, vol. 15, Dec. 2015.

[16] R. Patel, H. Shetty, R. Jackson, M. Broadbent, R. Stewart, J. Boydell, P. McGuire, and M. Taylor, "Delays before Diagnosis and Initiation of Treatment in Patients Presenting to Mental Health Services with Bipolar Disorder," *PLOS ONE*, vol. 10, p. e0126530, May 2015.

[17] R. Patel, N. Jayatilleke, R. Jackson, R. Stewart, and P. McGuire, "Investigation of negative symptoms in schizophrenia with a machine learning text-mining approach," *The Lancet*, vol. 383, p. S16, Feb. 2014.

[18] A. C. Fernandes, D. Cloete, M. T. Broadbent, R. D. Hayes, C.-K. Chang, R. G. Jackson, A. Roberts, J. Tsang, M. Soncul, J. Liebscher, R. Stewart, and F. Callard, "Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records," *BMC Medical Informatics and Decision Making*, vol. 13, no. 1, p. 71, 2013.

[19] C.-Y. Wu, C.-K. Chang, D. Robson, R. Jackson, S.-J. Chen, R. D. Hayes, and R. Stewart, "Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register," *PloS one*, vol. 8, no. 9, p. e74262, 2013.

[20] G. Gorrell, R. Jackson, and A. Roberts, "Finding Negative Symptoms of Schizophrenia in Patient Records," in *Proc NLP Med Biol Work (NLPMedBio)*, (Hissar, Bulgaria), pp. 9–17, 2013.

[21] R. M. Gardner, J. M. Overhage, E. B. Steen, B. S. Munger, J. H. Holmes, J. J. Williamson, D. E. Detmer, and AMIA Board of Directors, "Core Content for the Subspecialty of Clinical Informatics," *Journal of the American Medical Informatics Association*, vol. 16, pp. 153–157, Dec. 2008. Cited by 0000.

[22] "ISO/TR 20514:2005(en) Health informatics — Electronic health record — Definition, scope and context."

[23] W. Phillips, K. Gorwitz, and A. K. Bahn, "Electronic maintenance of case registers," *Public Health Reports*, vol. 77, pp. 503–510, June 1962.

[24] E. M. Laska, "29. The Multi-State Information System for Psychiatric patients," *Medical care*, vol. 14, pp. 223–229, May 1976.

[25] W. J. Curran, E. M. Laska, H. Kaplan, and R. Bank, "Protection of privacy and confidentiality," *Science (New York, N.Y.)*, vol. 182, pp. 797–802, Nov. 1973. Cited by 0036.

[26] A. Takian, D. Petrakaki, T. Cornford, A. Sheikh, N. Barber, and National NHS Care Records Service Evaluation Team, "Building a house on shifting sand: Methodological considerations when evaluating the implementation and adoption of national electronic health record systems," *BMC health services research*, vol. 12, p. 105, 2012.

[27] R. Cornet and N. de Keizer, "Forty years of SNOMED: A literature review," *BMC Medical Informatics and Decision Making*, vol. 8, no. Suppl 1, p. S2, 2008.

[28] D. W. Simborg, "An emerging standard for health communications: The HL7 standard," *Healthcare Computing & Communications*, vol. 4, pp. 58, 60, Oct. 1987.

[29] W. H. Organization, *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. Geneva: World Health Organization, 1993.

[30] D. Lee, R. Cornet, F. Lau, and N. de Keizer, "A survey of SNOMED CT implementations," *Journal of Biomedical Informatics*, vol. 46, pp. 87–96, Feb. 2013.

[31] M. Barnes, "Lessons learned from the implementation of clinical messaging systems," *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pp. 36–40, 2007.

[32] "The Future of Healthcare Informatics: It Is Not What You Think," *Global Advances in Health and Medicine*, vol. 1, pp. 5–6, Sept. 2012.

[33] D. Gordon, "Merging multiple institutions: Information architecture problems and solutions," *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pp. 785–789, 1999.

[34] V. N. Hicken, S. N. Thornton, and R. A. Rocha, "Integration challenges of clinical information systems developed without a shared data dictionary," *Studies in Health Technology and Informatics*, vol. 107, no. Pt 2, pp. 1053–1057, 2004.

[35] I. M. Xierali, C.-J. Hsiao, J. C. Puffer, L. A. Green, J. C. B. Rinaldo, A. W. Bazemore, M. T. Burke, and R. L. Phillips, Jr, "The rise of electronic health record adop-

tion among family physicians," *Annals of family medicine*, vol. 11, no. 1, pp. 14–19, 2013 Jan-Feb.

[36] T. H. Payne, D. E. Detmer, J. C. Wyatt, and I. E. Buchan, "National-scale clinical information exchange in the United Kingdom: Lessons for the United States," *Journal of the American Medical Informatics Association*, vol. 18, pp. 91–98, Jan. 2011.

[37] A. Clarke, J. Adamson, L. Sheard, P. Cairns, I. Watt, and J. Wright, "Implementing electronic patient record systems (EPRs) into England's acute, mental health and community care trusts: A mixed methods study," *BMC Medical Informatics and Decision Making*, vol. 15, Dec. 2015.

[38] S. W. T. F. null, "NHS England. Safer hospitals safer wards achieving an integrated digital care record. Guidance and launch of the Safer Hospitals," 2013.

[39] G. C. Liu, J. G. Cooper, K. M. Schoeffler, and W. E. Hammond, "Standards for the electronic health record, emerging from health care's Tower of Babel," *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pp. 388–392, 2001.

[40] S. M. Retchin and R. P. Wenzel, "Electronic medical record systems at academic health centers: Advantages and implementation issues," *Academic medicine: journal of the Association of American Medical Colleges*, vol. 74, pp. 493–498, May 1999.

[41] B. Blobel, "Comparing concepts for electronic health record architectures," *Studies in health technology and informatics*, vol. 90, pp. 209–214, 2002. Cited by 0009.

[42] L. Poissant, J. Pereira, R. Tamblyn, and Y. Kawasumi, "The impact of electronic health records on time efficiency of physicians and nurses: A systematic review," *Journal of the American Medical Informatics Association: JAMIA*, vol. 12, no. 5, pp. 505–516, 2005 Sep-Oct.

[43] W. J. van der Kam, B. Meyboom de Jong, T. F. Tromp, P. W. Moorman, and J. van der Lei, "Effects of electronic communication between the GP and the pharmacist. The quality of medication data on admission and after discharge," *Family Practice*, vol. 18, pp. 605–609, Dec. 2001.

[44] W. Rosenberg and A. Donald, "Evidence based medicine: An approach to clinical problem-solving," *BMJ*, vol. 310, pp. 1122–1126, Apr. 1995.

[45] M. Clarke, "The 1944 patulin trial of the British Medical Research Council," *Journal of the Royal Society of Medicine*, vol. 99, pp. 478–480, Sept. 2006.

[46] A. L. Cochrane, *Effectiveness and Efficiency: Random Reflections on Health Services ; the Rock Carling Fellowship 1971.* London: British Medical Journal [u.a.], reprinted ed., 1999. OCLC: 246498726.

[47] D. L. Sackett, W. M. C. Rosenberg, J. A. M. Gray, R. B. Haynes, and W. S. Richardson, "Evidence based medicine: What it is and what it isn't," *BMJ*, vol. 312, pp. 71–72, Jan. 1996.

[48] Evidence-Based Medicine Working Group, "Evidence-based medicine. A new approach to teaching the practice of medicine," *JAMA*, vol. 268, pp. 2420–2425, Nov. 1992.

[49] D. M. Eddy, "Practice policies: Where do they come from?," *JAMA*, vol. 263, pp. 1265, 1269, 1272 passim, Mar. 1990.

[50] D. Isaacs and D. Fitzgerald, "Seven alternatives to evidence based medicine," *BMJ (Clinical research ed.)*, vol. 319, no. 7225, p. 1618, 1999 Dec 18-25.

[51] H. C. Sox, M. Helfand, J. Grimshaw, K. Dickersin, D. Tovey, J. A. Knottnerus, and P. Tugwell, "Comparative Effectiveness Research: Challenges for Medical Journals," *Journal of Clinical Epidemiology*, vol. 63, pp. 862–864, Aug. 2010.

[52] L. Smeeth, A. J. Hall, E. Fombonne, L. C. Rodrigues, X. Huang, and P. G. Smith, "A case-control study of autism and mumps-measles-rubella vaccination using the general practice research database: Design and methodology," *BMC public health*, vol. 1, p. 2, 2001.

[53] J. L. Annest, J. A. Mercy, D. R. Gibson, and G. W. Ryan, "National estimates of nonfatal firearm-related injuries. Beyond the tip of the iceberg," *JAMA*, vol. 273, pp. 1749–1754, June 1995.

[54] R. Haux, "Health information systems - past, present, future," *International journal of medical informatics*, vol. 75, no. 3-4, pp. 268–281, 2006 Mar-Apr. Cited by 0365.

[55] A. J. Kriechbaum and M. G. Baker, "The epidemiology of imported malaria in New Zealand 1980-92," *The New Zealand Medical Journal*, vol. 109, pp. 405–407, Oct. 1996.

[56] W. M. Tierney, B. Y. Takesue, D. L. Vargo, and X. H. Zhou, "Using electronic medical records to predict mortality in primary care patients with heart disease: Prognostic power and pathophysiologic implications," *Journal of General Internal Medicine*, vol. 11, pp. 83–91, Feb. 1996.

[57] R. Haux, E. Ammenwerth, W. Herzog, and P. Knaup, "Health care in the information society. A prognosis for the year 2013," *International Journal of Medical Informatics*, vol. 66, pp. 3–21, Nov. 2002. Cited by 0196.

[58] A. Geissbuhler, C. Safran, I. Buchan, R. Bellazzi, S. Labkoff, K. Eilenberg, A. Leese, C. Richardson, J. Mantas, P. Murray, and G. De Moor, "Trustworthy reuse of health data: A transnational perspective," *International Journal of Medical Informatics*, vol. 82, pp. 1–9, Jan. 2013.

[59] F. Godlee, "What can we salvage from care.data?," *BMJ*, p. i3907, July 2016.

[60] T.-P. van Staa, B. Goldacre, I. Buchan, and L. Smeeth, "Big health data: The need to earn public trust," *BMJ*, p. i3636, July 2016.

[61] N. C. Lea, J. Nicholls, C. Dobbs, N. Sethi, J. Cunningham, J. Ainsworth, M. Heaven, T. Peacock, A. Peacock, K. Jones, G. Laurie, and D. Kalra, "Data Safe Havens and Trust: Toward a Common Understanding of Trusted Research Platforms for Governing Secure and Ethical Health Research," *JMIR Medical Informatics*, vol. 4, p. e22, June 2016.

[62] P. R. Burton, M. J. Murtagh, A. Boyd, J. B. Williams, E. S. Dove, S. E. Wallace, A.-M. Tassé, J. Little, R. L. Chisholm, A. Gaye, K. Hveem, A. J. Brookes, P. Goodwin, J. Fistein, M. Bobrow, and B. M. Knoppers, "Data Safe Havens in health research and healthcare," *Bioinformatics*, vol. 31, pp. 3241–3248, Oct. 2015.

[63] R. Stewart and K. Davis, "'Big data' in mental health research: Current status and emerging possibilities," *Social Psychiatry and Psychiatric Epidemiology*, vol. 51, pp. 1055–1072, Aug. 2016.

[64] R. Stewart, M. Soremekun, G. Perera, M. Broadbent, F. Callard, M. Denis, M. Hotopf, G. Thornicroft, and S. Lovestone, "The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: Development and descriptive data," *BMC psychiatry*, vol. 9, p. 51, 2009. Cited by 0028.

[65] K. H. Jones, D. V. Ford, C. Jones, R. Dsilva, S. Thompson, C. J. Brooks, M. L. Heaven, D. S. Thayer, C. L. McNerney, and R. A. Lyons, "A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: A privacy-protecting remote access system for health-related research and evaluation," *Journal of Biomedical Informatics*, vol. 50, pp. 196–204, Aug. 2014.

[66] "Data security incident trends." `https://ico.org.uk/action-weve-taken/data-security-incident-trends/`. Accessed: 2016-09-30.

[67] B. Sadan, "Patient data confidentiality and patient rights," *International Journal of Medical Informatics*, vol. 62, pp. 41–49, June 2001.

[68] S. D. Persell, J. M. Wright, J. A. Thompson, K. S. Kmetik, and D. W. Baker, "Assessing the validity of national quality measures for coronary artery disease using an electronic health record," *Archives of internal medicine*, vol. 166, pp. 2272–2277, Nov. 2006. Cited by 0061.

[69] C. P. Friedman and U. L. Abbas, "Is medical informatics a mature science? A review of measurement practice in outcome studies of clinical systems," *International journal of medical informatics*, vol. 69, pp. 261–272, Mar. 2003. Cited by 0032.

[70] J. Wyatt, "Medical informatics, artefacts or science?," *Methods of information in medicine*, vol. 35, pp. 197–200, Sept. 1996.

[71] P. Munk-Jørgensen, N. Okkels, D. Golberg, M. Ruggeri, and G. Thornicroft, "Fifty years' development and future perspectives of psychiatric register research," *Acta Psychiatrica Scandinavica*, vol. 130, pp. 87–98, Aug. 2014.

[72] S. Tyree, A. Meyer, and D. SB, *Linking Data for Health Services Research: A Framework and Instructional Guide [Rockville].* Agency for Healthcare Research and Quality, 2014.

[73] E. Herrett, A. D. Shah, R. Boggon, S. Denaxas, L. Smeeth, T. van Staa, A. Timmis, and H. Hemingway, "Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: Cohort study," *BMJ*, vol. 346, pp. f2350–f2350, May 2013.

[74] P. H. R. D. F. null, "Enabling Data Linkage to Maximise the Value of Public Health Research Data," Mar. 2015.

[75] C. Wallace, P. Mullen, P. Burgess, S. Palmer, D. Ruschena, and C. Browne, "Serious criminal offending and mental disorder. Case linkage study," *The British Journal of Psychiatry*, vol. 172, pp. 477–484, June 1998.

[76] S. Fazel and M. Grann, "The Population Impact of Severe Mental Illness on Violent Crime," *American Journal of Psychiatry*, vol. 163, pp. 1397–1403, Aug. 2006.

[77] H. A. Herinckx, S. C. Swart, S. M. Ama, C. D. Dolezal, and S. King, "Rearrest and Linkage to Mental Health Services Among Clients of the Clark County Mental Health Court Program," *Psychiatric Services*, vol. 56, pp. 853–857, July 2005.

[78] T. Botsis, G. Hartvigsen, F. Chen, and C. Weng, "Secondary use of EHR: Data quality issues and informatics opportunities," *AMIA Summits Transl Sci Proc*, vol. 2010, pp. 1–5, 2010.

[79] K. S. Chan, J. B. Fowles, and J. P. Weiner, "Review: Electronic Health Records and the Reliability and Validity of Quality Measures: A Review of the Literature," *Medical Care Research and Review*, vol. 67, pp. 503–527, Oct. 2010.

[80] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research," *Journal of the American Medical Informatics Association*, vol. 20, pp. 144–151, Jan. 2013.

[81] T. Botsis, G. Hartvigsen, F. Chen, and C. Weng, "Secondary Use of EHR: Data Quality Issues and Informatics Opportunities," *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science*, vol. 2010, pp. 1–5, 2010.

[82] T. Benson, "The history of the Read Codes: The inaugural James Read Memorial Lecture 2011," *Informatics in Primary Care*, vol. 19, no. 3, pp. 173–182, 2011.

[83] E. Ford, J. A. Carroll, H. E. Smith, D. Scott, and J. A. Cassell, "Extracting information from the text of electronic medical records to improve case detection: A systematic review," *Journal of the American Medical Informatics Association*, vol. 23, pp. 1007–1015, Sept. 2016.

[84] T. Greenhalgh and B. Hurwitz, "Narrative based medicine: Why study narrative?," *BMJ*, vol. 318, pp. 48–50, Jan. 1999.

[85] H. D. Stein, P. Nadkarni, J. Erdos, and P. L. Miller, "Exploring the Degree of Concordance of Coded and Textual Data in Answering Clinical Queries from a Clinical Data Repository," *Journal of the American Medical Informatics Association*, vol. 7, pp. 42–54, Jan. 2000.

[86] E. Ford, A. Nicholson, R. Koeling, A. Tate, J. Carroll, L. Axelrod, H. E. Smith, G. Rait, K. A. Davies, I. Petersen, T. Williams, and J. A. Cassell, "Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: What information is hidden in free text?," *BMC medical research methodology*, vol. 13, p. 105, 2013.

[87] S. de Lusignan, S. E. Wells, N. J. Hague, and K. Thiru, "Managers see the problems associated with coding clinical data as a technical issue whilst clinicians also see

cultural barriers," *Methods of Information in Medicine*, vol. 42, no. 4, pp. 416–422, 2003.

[88] S. H. Walsh, "The clinician's perspective on electronic health records and how they can affect patient care," *BMJ*, vol. 328, pp. 1184–1187, May 2004.

[89] S. M. Powsner, J. C. Wyatt, and P. Wright, "Opportunities for and challenges of computerisation," *Lancet (London, England)*, vol. 352, pp. 1617–1622, Nov. 1998.

[90] G. Mikkelsen and J. Aasly, "Consequences of impaired data quality on information retrieval in electronic patient records," *International Journal of Medical Informatics*, vol. 74, pp. 387–394, June 2005.

[91] S. Lewis, "Brave new EMR," *Annals of internal medicine*, vol. 154, pp. 368–369, Mar. 2011.

[92] M. Hilbert and P. Lopez, "The World's Technological Capacity to Store, Communicate, and Compute Information," *Science*, vol. 332, pp. 60–65, Apr. 2011.

[93] A. Pelzer, G. Tibaux, and R. Chef, "[Experience with the computer processing of cytologic reports in natural language]," *Minerva Ginecologica*, vol. 22, pp. 1175–1179, Dec. 1970.

[94] C. Shivade, P. Raghavan, E. Fosler-Lussier, P. J. Embi, N. Elhadad, S. B. Johnson, and A. M. Lai, "A review of approaches to identifying patient phenotype cohorts using electronic health records," *Journal of the American Medical Informatics Association*, vol. 21, pp. 221–230, Nov. 2013.

[95] H. Cunningham, V. Tablan, A. Roberts, and K. Bontcheva, "Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics," *PLoS Computational Biology*, vol. 9, p. e1002854, Feb. 2013.

[96] D. Ferrucci and A. Lally, "UIMA: An architectural approach to unstructured information processing in the corporate research environment," *Natural Language Engineering*, vol. 10, pp. 327–348, Sept. 2004.

[97] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. Beijing ; Cambridge [Mass.]: O'Reilly, 1st ed ed., 2009. OCLC: ocn301885973.

[98] Z. S. Harris, *A Grammar of English on Mathematical Principles*. John Wiley & Sons Inc, 1982.

[99] Z. Harris, "Theory of language and information: A mathematical approach," 1991.

[100] C. Friedman, P. Kra, and A. Rzhetsky, "Two biomedical sublanguages: A description based on the theories of Zellig Harris," *Journal of Biomedical Informatics*, vol. 35, pp. 222–235, Aug. 2002.

[101] A. Akkasi, E. Varoğlu, and N. Dimililer, "ChemTok: A New Rule Based Tokenizer for Chemical Named Entity Recognition," *BioMed Research International*, vol. 2016, p. 4248026, 2016.

[102] H. Chen, B. Martin, C. M. Daimon, S. Siddiqui, L. M. Luttrell, and S. Maudsley, "Textrous!: Extracting semantic textual meaning from gene sets," *PloS One*, vol. 8, no. 4, p. e62665, 2013.

[103] I. Merriam-Webster, *The Merriam-Webster Dictionary.* 2016. OCLC: 921867715.

[104] D. Rudrapal, A. Jamatia, K. Chakma, A. Das, and B. Gambäck, "Sentence boundary detection for social media text," in *12th International Conference on Natural Language Processing, At IIITM-Kerala, Trivandrum, Volume 12*, 12 2015.

[105] D. Griffis, C. Shivade, E. Fosler-Lussier, and A. M. Lai, "A Quantitative and Qualitative Evaluation of Sentence Boundary Detection for the Clinical Domain," *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, vol. 2016, pp. 88–97, 2016.

[106] R. Leaman, R. Khare, and Z. Lu, "Challenges in clinical natural language processing for automated disorder normalization," *Journal of Biomedical Informatics*, vol. 57, pp. 28–37, Oct. 2015.

[107] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10, p. 707, Feb. 1966.

[108] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Communications of the ACM*, vol. 7, pp. 171–176, Mar. 1964.

[109] W. E. Winkler, "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," in *Proceedings of the Section on Survey Research*, (Wachington, DC), pp. 354–359, 1990.

[110] J. Zheng, W. W. Chapman, R. S. Crowley, and G. K. Savova, "Coreference resolution: A review of general methodologies and applications in the clinical domain," *Journal of Biomedical Informatics*, vol. 44, pp. 1113–1122, Dec. 2011.

[111] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[112] G. Minnen, J. Carroll, and D. Pearce, "Applied morphological processing of English," *Natural Language Engineering*, vol. 7, no. 03, pp. 207–223, 2001.

[113] S. B. null, *Part-of-Speech Tagging Guidelines for the Penn Treebank Project*. University of Pennsylvania, 1990.

[114] W. Ling, T. Lúis, L. Marujo, R. F. Astudillo, S. Amir, C. Dyer, A. W. Black, and I. Trancoso, "Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation," *CoRR*, vol. abs/1508.02096, 2015.

[115] U. Hahn and J. Wermter, "High-performance tagging on medical texts," in *Proceedings of the 20th International Conference on Computational Linguistics*, p. 973, Association for Computational Linguistics, 2004.

[116] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, pp. 507–513, Sept. 2010.

[117] J.-w. Fan, R. Prasad, R. M. Yabut, R. M. Loomis, D. S. Zisook, J. E. Mattison, and Y. Huang, "Part-of-speech tagging for clinical text: Wall or bridge between institutions?," *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, vol. 2011, pp. 382–391, 2011.

[118] L. A. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning," in *Natural language processing using very large corpora*, pp. 157–176, Springer, 1999.

[119] A. Savkov, J. Carroll, and J. Cassell, "Chunking clinical text containing non-canonical language," in *Proceedings of BioNLP*, 2014.

[120] D. Andor, C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins, "Globally Normalized Transition-Based Neural Networks," *ArXiv e-prints*, Mar. 2016.

[121] M. Jiang, Y. Huang, J.-w. Fan, B. Tang, J. Denny, and H. Xu, "Parsing clinical text: How good are the state-of-the-art parsers?," *BMC Medical Informatics and Decision Making*, vol. 15, no. Suppl 1, p. S2, 2015.

[122] S. Velupillai, D. Mowery, B. R. South, M. Kvist, and H. Dalianis, "Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis:," *IMIA Yearbook*, vol. 10, no. 1, pp. 183–193, 2015.

[123] N. Elhadad, S. Pradhan, W. Chapman, S. Manandhar, and G. Savova, "SemEval-2015 task 14: Analysis of clinical text," in *Proc of Workshop on Semantic Evaluation. Association for Computational Linguistics*, pp. 303–10, 2015.

[124] H. Liu, A. R. Aronson, and C. Friedman, "A study of abbreviations in MEDLINE abstracts," *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pp. 464–468, 2002.

[125] Y. Wu, J. C. Denny, S. T. Rosenbloom, R. A. Miller, D. A. Giuse, L. Wang, C. Blanquicett, E. Soysal, J. Xu, and H. Xu, "A long journey to short abbreviations: Developing an open-source framework for clinical abbreviation recognition and disambiguation (CARD)," *Journal of the American Medical Informatics Association*, p. ocw109, Aug. 2016.

[126] W. Sun, A. Rumshisky, and O. Uzuner, "Temporal reasoning over clinical text: The state of the art," *Journal of the American Medical Informatics Association*, vol. 20, pp. 814–819, May 2013.

[127] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, "A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries," *Journal of Biomedical Informatics*, vol. 34, pp. 301–310, Oct. 2001.

[128] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle, "Extracting information from textual documents in the electronic health record: A review of recent research," *Yearbook of medical informatics*, pp. 128–144, 2008. Cited by 0180.

[129] T. Groza, S. Köhler, D. Moldenhauer, N. Vasilevsky, G. Baynam, T. Zemojtel, L. M. Schriml, W. A. Kibbe, P. N. Schofield, T. Beck, D. Vasant, A. J. Brookes, A. Zankl, N. L. Washington, C. J. Mungall, S. E. Lewis, M. A. Haendel, H. Parkinson, and P. N. Robinson, "The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease," *The American Journal of Human Genetics*, vol. 97, pp. 111–124, July 2015.

[130] H. Harkema, J. N. Dowling, T. Thornblade, and W. W. Chapman, "ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports," *Journal of biomedical informatics*, vol. 42, no. 5, pp. 839–851, 2009.

[131] H. Harkema, J. N. Dowling, T. Thornblade, and W. W. Chapman, "ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports," *Journal of Biomedical Informatics*, vol. 42, pp. 839–851, Oct. 2009.

[132] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii, "BRAT: A web-based tool for NLP-assisted text annotation," in *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 102–107, Association for Computational Linguistics, 2012.

[133] P. V. Ogren, "Knowtator: A protégé plug-in for annotated corpus construction," in *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume: Demonstrations*, pp. 273–275, Association for Computational Linguistics, 2006.

[134] H. Kucera and W. Francis, *A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers (Revised and Amplified from 1967 Version)*. Providence, RI: Brown University Press, 1979.

[135] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of English: The Penn Treebank," *Computational linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[136] A. Taylor, M. Marcus, and B. Santorini, "The Penn Treebank: An Overview," in *Treebanks* (N. Ide, J. Véronis, and A. Abeillé, eds.), vol. 20, pp. 5–22, Dordrecht: Springer Netherlands, 2003.

[137] H. Cunningham, H. Cunningham, D. Maynard, D. Maynard, V. Tablan, and V. Tablan, "JAPE: A Java Annotation Patterns Engine," 1999.

[138] P. Kluegl, M. Toepfer, P.-D. Beck, G. Fette, and F. Puppe, "UIMA Ruta: Rapid development of rule-based information extraction applications," *Natural Language Engineering*, vol. 22, no. 01, pp. 1–40, 2016.

[139] J. Strötgen and M. Gertz, "HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, (Los Angeles, California), pp. 321–324, Association for Computational Linguistics, 2010.

[140] H. Suominen, F. Ginter, S. Pyysalo, A. Airola, T. Pahikkala, S. Salanter, and T. Salakoski, "Machine learning to automate the assignment of diagnosis codes to free-text radiology reports: A method description," in *Proceedings of the ICML/UAI/COLT Workshop on Machine Learning for Health-Care Applications*, 2008.

[141] W. W. Cohen, "Fast effective rule induction," in *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 115–123, 1995.

[142] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: An introduction," *Journal of the American Medical Informatics Association*, vol. 18, pp. 544–551, Sept. 2011.

[143] L. Chiticariu, Y. Li, and F. R. Reiss, "Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems!," in *EMNLP*, pp. 827–832, 2013.

[144] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, vol. 7, pp. 179–188, Sept. 1936.

[145] D. Koller and S. Tong, "Support vector machine active learning with applications to text classification," *The Journal of Machine Learning Research*, vol. 2, pp. 45–66, 2001.

[146] D. Li, K. Kipper-Schuler, and G. Savova, "Conditional Random Fields and Support Vector Machines for Disorder Named Entity Recognition in Clinical Texts," in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, (Columbus, Ohio), pp. 94–95, Association for Computational Linguistics, 2008.

[147] Z. Afzal, M. J. Schuemie, J. C. van Blijderveen, E. F. Sen, M. C. J. M. Sturkenboom, and J. A. Kors, "Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records," *BMC medical informatics and decision making*, vol. 13, p. 30, 2013.

[148] R. Khor, W.-K. Yip, M. Bressel, W. Rose, G. Duchesne, and F. Foroudi, "Practical implementation of an existing smoking detection pipeline and reduced support vector machine training corpus requirements," *Journal of the American Medical Informatics Association*, vol. 21, pp. 27–30, Aug. 2013.

[149] J. P. Ferraro, H. Daumé, S. L. DuVall, W. W. Chapman, H. Harkema, and P. J. Haug, "Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation," *Journal of the American Medical Informatics Association*, vol. 20, pp. 931–939, Sept. 2013.

[150] W. W. Chapman, P. M. Nadkarni, L. Hirschman, L. W. D'Avolio, G. K. Savova, and O. Uzuner, "Overcoming barriers to NLP for clinical text: The role of shared tasks

and the need for additional creative solutions," *Journal of the American Medical Informatics Association*, vol. 18, pp. 540–543, Sept. 2011.

[151] F. Xia and M. Yetisgen-Yildiz, "Clinical corpus annotation: Challenges and strategies," in *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM'2012) in Conjunction with the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey*, 2012.

[152] A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, I. Roberts, and A. Setzer, "Building a semantically annotated corpus of clinical texts," *Journal of Biomedical Informatics*, vol. 42, pp. 950–966, Oct. 2009.

[153] Ö. Uzuner, I. Solti, and E. Cadag, "Extracting medication information from clinical text," *Journal of the American Medical Informatics Association*, vol. 17, pp. 514–518, Sept. 2010.

[154] J. Patrick and M. Li, "High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge," *Journal of the American Medical Informatics Association*, vol. 17, pp. 524–527, Sept. 2010.

[155] H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny, "MedEx: A medication information extraction system for clinical narratives," *Journal of the American Medical Informatics Association*, vol. 17, pp. 19–24, Jan. 2010.

[156] I. Spasić, F. Sarafraz, J. A. Keane, and G. Nenadić, "Medication information extraction with linguistic pattern matching and semantic rules," *Journal of the American Medical Informatics Association*, vol. 17, pp. 532–535, Sept. 2010.

[157] A. Stubbs, C. Kotfila, H. Xu, and Ö. Uzuner, "Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2," *Journal of Biomedical Informatics*, vol. 58, pp. S67–S77, Dec. 2015.

[158] K. Roberts, S. E. Shooshan, L. Rodriguez, S. Abhyankar, H. Kilicoglu, and D. Demner-Fushman, "The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs," *Journal of Biomedical Informatics*, vol. 58, pp. S111–S119, Dec. 2015.

[159] M. Torii, J.-w. Fan, W.-l. Yang, T. Lee, M. T. Wiley, D. S. Zisook, and Y. Huang, "Risk factor detection for heart disease by applying text analytics in electronic medical records," *Journal of Biomedical Informatics*, vol. 58, pp. S164–S170, Dec. 2015.

[160] H. Yang and J. M. Garibaldi, "A hybrid model for automatic identification of risk factors for heart disease," *Journal of Biomedical Informatics*, vol. 58, pp. S171–S182, Dec. 2015.

[161] A. Stubbs, M. Filannino, and Ö. Uzuner, "De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks Track 1," *Journal of Biomedical Informatics*, vol. 75, pp. S4–S18, Nov. 2017.

[162] Z. Liu, B. Tang, X. Wang, and Q. Chen, "De-identification of clinical notes via recurrent neural network and conditional random field," *Journal of Biomedical Informatics*, vol. 75, pp. S34–S42, Nov. 2017.

[163] M. Filannino, A. Stubbs, and Ö. Uzuner, "Symptom severity prediction from neuropsychiatric clinical records: Overview of 2016 CEGS N-GRID shared tasks Track 2," *Journal of Biomedical Informatics*, vol. 75, pp. S62–S70, Nov. 2017.

[164] T. Tran and R. Kavuluru, "Predicting mental conditions based on "history of present illness" in psychiatric notes with deep neural networks," *Journal of Biomedical Informatics*, vol. 75, pp. S138–S148, Nov. 2017.

[165] H.-J. Dai, E. C.-Y. Su, M. Uddin, J. Jonnagaddala, C.-S. Wu, and S. Syed-Abdul, "Exploring associations of clinical and social parameters with violent behaviors among psychiatric patients," *Journal of Biomedical Informatics*, vol. 75, pp. S149–S159, Nov. 2017.

[166] Y. Zhang, O. Zhang, Y. Wu, H.-J. Lee, J. Xu, H. Xu, and K. Roberts, "Psychiatric symptom recognition without labeled data using distributional representations of phrases and on-line knowledge," *Journal of Biomedical Informatics*, vol. 75, pp. S129–S137, Nov. 2017.

[167] E. M. Voorhees and W. R. Hersh, "Overview of the TREC 2012 Medical Records Track.," in *TREC*, 2012.

[168] S. Bethard, L. Derczynski, G. Savova, G. Savova, J. Pustejovsky, and M. Verhagen, "Semeval-2015 task 6: Clinical tempeval," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 806–814, 2015.

[169] A. Rector, J. Rogers, A. Taweel, D. Ingram, D. Kalra, J. Milan, P. Singleton, R. Gaizauskas, M. Hepple, D. Scott, and others, "Clef: Joining up healthcare with clinical and post-genomic research," 2003.

[170] N. Alnazzawi, P. Thompson, R. Batista-Navarro, and S. Ananiadou, "Using text mining techniques to extract phenotypic information from the PhenoCHF corpus," *BMC Medical Informatics and Decision Making*, vol. 15, Dec. 2015.

[171] N. Alnazzawi, P. Thompson, and S. Ananiadou, "Mapping Phenotypic Information in Heterogeneous Textual Sources to a Domain-Specific Terminological Resource," *PLOS ONE*, vol. 11, p. e0162287, Sept. 2016.

[172] N. Alnazzawi, P. Thompson, and S. Ananiadou, "Building a semantically annotated corpus for congestive heart and renal failure from clinical records and the literature," in *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, pp. 69–74, 2014.

[173] P. Przybyła, M. Shardlow, S. Aubin, R. Bossy, R. Eckart de Castilho, S. Piperidis, J. McNaught, and S. Ananiadou, "Text mining resources for the life sciences," *Database*, vol. 2016, 2016.

[174] C. Friedman, P. O. Alderson, J. H. M. Austin, J. J. Cimino, and S. B. Johnson, "A General Natural-language Text Processor for Clinical Radiology," *Journal of the American Medical Informatics Association*, vol. 1, pp. 161–174, Mar. 1994.

[175] C. Friedman, "A broad-coverage natural language processing system," *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pp. 270–274, 2000.

[176] S. Bakken, S. Hyun, C. Friedman, and S. Johnson, "A comparison of semantic categories of the ISO reference terminology models for nursing and the MedLEE natural language processing system," *Stud Health Technol Inform*, vol. 107, no. Pt 1, pp. 472–476, 2004.

[177] J.-H. Chiang, J.-W. Lin, and C.-W. Yang, "Automated evaluation of electronic discharge notes to assess quality of care for cardiovascular diseases using Medical Language Extraction and Encoding System (MedLEE)," *Journal of the American Medical Informatics Association*, vol. 17, pp. 245–252, May 2010.

[178] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program," *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pp. 17–21, 2001.

[179] S. Pradhan, N. Elhadad, B. R. South, D. Martinez, L. Christensen, A. Vogel, H. Suominen, W. W. Chapman, and G. Savova, "Evaluating the state of the art in disorder recognition and normalization of the clinical narrative," *Journal of the American Medical Informatics Association*, Aug. 2014.

[180] S. Sohn and G. Savova, "Mayo clinic smoking status classification system: Extensions and improvements.," in *AMIA... Annual Symposium Proceedings/AMIA Symposium. AMIA Symposium*, vol. 2009, pp. 619–623, 2008.

[181] C. Lin, D. Dligach, T. A. Miller, S. Bethard, and G. K. Savova, "Multilayered temporal modeling for the clinical domain," *Journal of the American Medical Informatics Association*, vol. 23, pp. 387–395, Mar. 2016.

[182] R. Khor, W.-K. Yip, M. Bressel, W. Rose, G. Duchesne, and F. Foroudi, "Practical implementation of an existing smoking detection pipeline and reduced support vector machine training corpus requirements," *Journal of the American Medical Informatics Association*, vol. 21, pp. 27–30, Jan. 2014.

[183] G. Divita, Q. T. Zeng, A. V. Gundlapalli, S. Duvall, J. Nebeker, and M. H. Samore, "Sophia: A Expedient UMLS Concept Extraction Annotator," *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, vol. 2014, pp. 467–476, 2014.

[184] D. J. Cronkite and D. S. Carrell, "A Lightweight Text Mining Tool for Multisite Research," *Journal of Patient-Centered Research and Reviews*, vol. 2, p. 119, Apr. 2015.

[185] Y. Wu, J. C. Denny, S. Rosenbloom, R. A. Miller, D. A. Giuse, and H. Xu, "A comparative study of current Clinical Natural Language Processing systems on handling abbreviations in discharge summaries.," in *AMIA*, 2012.

[186] Y. Kano, W. A. Baumgartner, L. McCrohon, S. Ananiadou, K. B. Cohen, L. Hunter, and J. Tsujii, "U-Compare: Share and compare text mining tools with UIMA," *Bioinformatics*, vol. 25, pp. 1997–1998, Aug. 2009.

[187] R. Rak, A. Rowley, W. Black, and S. Ananiadou, "Argo: An integrative, interactive, text mining-based workbench supporting curation," *Database*, vol. 2012, pp. bas010–bas010, Mar. 2012.

[188] H. Liu, S. T. Wu, D. Li, S. Jonnalagadda, S. Sohn, K. Wagholikar, P. J. Haug, S. M. Huff, and C. G. Chute, "Towards a semantic lexicon for clinical natural language processing," *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2012, pp. 568–576, 2012.

[189] H. Cunningham, "Information extraction, automatic," *Encyclopedia of language and linguistics,*, pp. 665–677, 2005.

[190] H. Cunningham, D. Maynard, and K. Bontcheva, "Tablan., V.(2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications," in *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*.

[191] H. Cunningham, "GATE, a general architecture for text engineering," *Computers and the Humanities*, vol. 36, no. 2, pp. 223–254, 2002.

[192] N. Sager, M. Lyman, C. Bucknall, N. Nhan, and L. J. Tick, "Natural Language Processing and the Representation of Clinical Data," *Journal of the American Medical Informatics Association*, vol. 1, pp. 142–160, Mar. 1994.

[193] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, J. F. Hurdle, and others, "Extracting information from textual documents in the electronic health record: A review of recent research," *Yearb Med Inform*, vol. 35, pp. 128–44, 2008.

[194] C.-Y. Wu, C.-K. Chang, D. Robson, R. Jackson, S.-J. Chen, R. D. Hayes, and R. Stewart, "Evaluation of Smoking Status Identification Using Electronic Health Records and Open-Text Information in a Large Mental Health Case Register," *PLoS ONE*, vol. 8, p. e74262, Sept. 2013.

[195] J. Sultana, C. K. Chang, R. D. Hayes, M. Broadbent, R. Stewart, A. Corbett, and C. Ballard, "Associations between risk of mortality and atypical antipsychotic use in vascular dementia: A clinical cohort study: Antipsychotics and mortality in vascular dementia," *International Journal of Geriatric Psychiatry*, vol. 29, pp. 1249–1254, Dec. 2014.

[196] Y.-P. Su, C.-K. Chang, R. D. Hayes, G. Perera, M. Broadbent, D. To, M. Hotopf, and R. Stewart, "Mini-Mental State Examination as a Predictor of Mortality among Older People Referred to Secondary Mental Healthcare," *PLoS ONE*, vol. 9, p. e105312, Sept. 2014.

[197] G. Perera, M. Khondoker, M. Broadbent, G. Breen, and R. Stewart, "Factors Associated with Response to Acetylcholinesterase Inhibition in Dementia: A Cohort Study from a Secondary Mental Health Care Case Register in London," *PLoS ONE*, vol. 9, p. e109484, Nov. 2014.

[198] L. Shi, L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. de Longueville, E. S. Kawasaki, K. Y. Lee, Y. Luo, Y. A. Sun, J. C. Willey, R. A. Setterquist, G. M. Fischer, W. Tong, Y. P. Dragan, D. J. Dix, F. W. Frueh, F. M. Goodsaid, D. Herman, R. V. Jensen, C. D. Johnson, E. K.

Lobenhofer, R. K. Puri, U. Scherf, J. Thierry-Mieg, C. Wang, M. Wilson, P. K. Wolber, L. Zhang, S. Amur, W. Bao, C. C. Barbacioru, A. B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X. M. Cao, T. A. Cebula, J. J. Chen, J. Cheng, T.-M. Chu, E. Chudin, J. Corson, J. C. Corton, L. J. Croner, C. Davies, T. S. Davison, G. Delenstarr, X. Deng, D. Dorris, A. C. Eklund, X.-h. Fan, H. Fang, S. Fulmer-Smentek, J. C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P. K. Haje, J. Han, T. Han, H. C. Harbottle, S. C. Harris, E. Hatchwell, C. A. Hauser, S. Hester, H. Hong, P. Hurban, S. A. Jackson, H. Ji, C. R. Knight, W. P. Kuo, J. E. LeClerc, S. Levy, Q.-Z. Li, C. Liu, Y. Liu, M. J. Lombardi, Y. Ma, S. R. Magnuson, B. Maqsodi, T. McDaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M. S. Orr, T. W. Osborn, A. Papallo, T. A. Patterson, R. G. Perkins, E. H. Peters, R. Peterson, K. L. Philips, P. S. Pine, L. Pusztai, F. Qian, H. Ren, M. Rosen, B. A. Rosenzweig, R. R. Samaha, M. Schena, G. P. Schroth, S. Shchegrova, D. D. Smith, F. Staedtler, Z. Su, H. Sun, Z. Szallasi, Z. Tezak, D. Thierry-Mieg, K. L. Thompson, I. Tikhonova, Y. Turpaz, B. Vallanat, C. Van, S. J. Walker, S. J. Wang, Y. Wang, R. Wolfinger, A. Wong, J. Wu, C. Xiao, Q. Xie, J. Xu, W. Yang, L. Zhang, S. Zhong, Y. Zong, and W. Slikker, "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements," *Nature Biotechnology*, vol. 24, pp. 1151–1161, Sept. 2006.

[199] Y. Zheng, T. Qing, Y. Song, J. Zhu, Y. Yu, W. Shi, L. Pusztai, and L. Shi, "Standardization efforts enabling next-generation sequencing and microarray based biomarkers for precision medicine," *Biomarkers in Medicine*, vol. 9, pp. 1265–1272, Nov. 2015.

[200] M. F. Wakelin and J. D. A. Widdowson, *Discovering English Dialects*. Shire Discovering classics, Princes Risborough: Shire, 4th ed ed., 2008.

[201] O. Patterson and J. F. Hurdle, "Document clustering of clinical narratives: A systematic study of clinical sublanguages," *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, vol. 2011, pp. 1099–1107, 2011.

[202] M. Kathy Giannangelo, C. RHIA, and others, "SNOMED CT survey: An assessment of implementation in EMR/EHR applications," *Perspectives in Health Information Management*, vol. 5, no. 7, p. 1, 2008.

[203] D. Lee, R. Cornet, F. Lau, and N. De Keizer, "A survey of SNOMED CT implementations," *Journal of biomedical informatics*, vol. 46, no. 1, pp. 87–96, 2013.

[204] A. L. Rector and others, "Clinical terminology: Why is it so hard?," *Methods of information in medicine*, vol. 38, no. 4/5, pp. 239–252, 1999.

[205] S. Wu, T. Miller, J. Masanz, M. Coarr, S. Halgrim, D. Carrell, and C. Clark, "Negation's not solved: Generalizability versus optimizability in clinical natural language processing," *PloS one*, vol. 9, no. 11, p. e112774, 2014.

[206] R. Koeling, A. R. Tate, and J. A. Carroll, "Automatically estimating the incidence of symptoms recorded in GP free text notes," in *Proceedings of the First International Workshop on Managing Interoperability and Complexity in Health Systems*, pp. 43–50, ACM, 2011.

[207] O. Zamaraeva, K. Howell, and A. Rhine, "Improving Feature Extraction for Pathology Reports with Precise Negation Scope Detection," in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3564–3575, 2018.

[208] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: An introduction," *Journal of the American Medical Informatics Association*, vol. 18, pp. 544–551, Sept. 2011.

[209] Ö. Uzuner, I. Solti, F. Xia, and E. Cadag, "Community annotation experiment for ground truth generation for the i2b2 medication challenge," *Journal of the American Medical Informatics Association*, vol. 17, pp. 519–523, Sept. 2010.

[210] J. Cormack, C. Nath, D. Milward, K. Raja, and S. R. Jonnalagadda, "Agile text mining for the 2014 i2b2/UTHealth Cardiac risk factors challenge," *Journal of Biomedical Informatics*, vol. 58, pp. S120–S127, Dec. 2015.

[211] K. Zheng, V. V. Vydiswaran, Y. Liu, Y. Wang, A. Stubbs, Ö. Uzuner, A. E. Gururaj, S. Bayer, J. Aberdeen, A. Rumshisky, S. Pakhomov, H. Liu, and H. Xu, "Ease of adoption of clinical natural language processing software: An evaluation of five systems," *Journal of Biomedical Informatics*, vol. 58, pp. S189–S196, Dec. 2015.

[212] "MetaMap Speed." `https://metamap.nlm.nih.gov/Docs/FAQ/Speed.pdf`. Accessed: 2016-09-14.

[213] D. Adam, "Mental health: On the spectrum," *Nature*, vol. 496, pp. 416–418, Apr. 2013.

[214] A. M. Kring, "The Clinical Assessment Interview for Negative Symptoms (CAINS): Final Development and Validation," *American Journal of Psychiatry*, vol. 170, p. 165, Feb. 2013. Cited by 0007.

[215] S. R. Kay, A. Fiszbein, and L. A. Opler, "The positive and negative syndrome scale (PANSS) for schizophrenia," *Schizophrenia bulletin*, vol. 13, no. 2, pp. 261–276, 1987. Cited by 8221.

[216] B. N. Axelrod, R. S. Goldman, and L. D. Alphs, "Validation of the 16-item Negative Symptom Assessment," *Journal of psychiatric research*, vol. 27, no. 3, pp. 253–258, 1993 Jul-Sep. Cited by 0016.

[217] E. Kiperwasser and Y. Goldberg, "Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations," *TACL*, vol. 4, pp. 313–327, 2016.

[218] Y. Zhang, F. Tiryaki, M. Jiang, and H. Xu, "Parsing clinical text: How good are the state-of-the-art deep learning based parsers?," in *2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W)*, pp. 80–81, IEEE, 2018.

[219] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *arXiv preprint arXiv:1508.05326*, 2015.

[220] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100, 000+ Questions for Machine Comprehension of Text," *CoRR*, vol. abs/1606.05250, 2016.

[221] A. Dumitrache, L. Aroyo, and C. Welty, "Crowdsourcing Ground Truth for Medical Relation Extraction," *CoRR*, vol. abs/1701.02185, 2017.

[222] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Machine learning*, vol. 15, no. 2, pp. 201–221, 1994.

[223] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *Journal of artificial intelligence research*, vol. 4, pp. 129–145, 1996.

[224] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *ICML*, vol. 2, p. 6, Citeseer, 2000.

[225] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Machine Learning Proceedings 1994*, pp. 148–156, Elsevier, 1994.

[226] A. Culotta and A. McCallum, "Reducing labeling effort for structured prediction tasks," in *AAAI*, vol. 5, pp. 746–751, 2005.

[227] K. Brinker, "Incorporating diversity in active learning with support vector machines," in *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 59–66, 2003.

[228] D. Shen, J. Zhang, J. Su, G. Zhou, and C. L. Tan, "Multi-Criteria-based Active Learning for Named Entity Recognition," in *ACL*, 2004.

[229] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, "Accelerating the Annotation of Sparse Named Entities by Dynamic Sentence Selection," in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, (Columbus, Ohio), pp. 30–37, Association for Computational Linguistics, June 2008.

[230] R. L. Figueroa, Q. Zeng-Treitler, L. H. Ngo, S. Goryachev, and E. P. Wiechmann, "Active learning for clinical text classification: Is it better than random sampling?," *Journal of the American Medical Informatics Association*, vol. 19, pp. 809–816, Sept. 2012.

[231] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu, 2002.

[232] J. Howard and S. Ruder, "Fine-tuned language models for text classification," *arXiv preprint arXiv:1801.06146*, 2018.

[233] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[234] A. Kolliakou, M. Ball, L. Derczynski, D. Chandran, G. Gkotsis, P. Deluca, R. Jackson, H. Shetty, and R. Stewart, "Novel psychoactive substances: An investigation of temporal trends in social media and electronic health records," *European Psychiatry*, vol. 38, pp. 15–21, 2016.

[235] R. Patel, *Investigating Clinical Outcomes in Psychotic Disorders Using an Electronic Case Register*. PhD thesis, King's College London, 2016.

[236] R. Schaefer, M. Broadbent, and M. Bruce, "Violent typologies among women inpatients with severe mental illness," *Social Psychiatry and Psychiatric Epidemiology*, vol. 51, pp. 1615–1622, Dec. 2016.

[237] Y. Kovalchuk, R. Stewart, M. Broadbent, T. J. P. Hubbard, and R. J. B. Dobson, "Analysis of diagnoses extracted from electronic health records in a large mental health case register," *PLOS ONE*, vol. 12, p. e0171526, Feb. 2017.

[238] J. Downs, H. Dean, S. Lechler, N. Sears, R. Patel, H. Shetty, M. Hotopf, T. Ford, M. Kyriakopoulos, C. M. Diaz-Caneja, C. Arango, J. H. MacCabe, R. D. Hayes, and L. Pina-Camacho, "Negative Symptoms in Early-Onset Psychosis and Their

Association With Antipsychotic Treatment Failure," *Schizophrenia Bulletin*, Jan. 2018.

[239] T. B. Forbush, A. V. Gundlapalli, M. N. Palmer, S. Shen, B. R. South, G. Divita, M. Carter, A. Redd, J. M. Butler, and M. Samore, ""Sitting on pins and needles": Characterization of symptom descriptions in clinical notes"," *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, vol. 2013, pp. 67–71, 2013.

[240] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *CoRR*, vol. abs/1802.05365, 2018.

[241] O. Uzuner, P. Szolovits, and I. Kohane, "I2b2 workshop on natural language processing challenges for clinical records," in *Proceedings of the Fall Symposium of the American Medical Informatics Association*, Citeseer, 2006.

[242] L. Deleger, Q. Li, T. Lingren, M. Kaiser, K. Molnar, L. Stoutenborough, M. Kouril, K. Marsolo, and I. Solti, "Building gold standard corpora for medical natural language processing tasks," *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2012, pp. 144–153, 2012.

[243] E. Tseytlin, K. Mitchell, E. Legowski, J. Corrigan, G. Chavan, and R. S. Jacobson, "NOBLE – Flexible concept recognition for large-scale biomedical natural language processing," *BMC Bioinformatics*, vol. 17, Dec. 2016.

[244] N. Alnazzawi, P. Thompson, and S. Ananiadou, "Mapping Phenotypic Information in Heterogeneous Textual Sources to a Domain-Specific Terminological Resource," *PLOS ONE*, vol. 11, p. e0162287, Sept. 2016.

[245] N. Siva, "UK gears up to decode 100 000 genomes from NHS patients," *The Lancet*, vol. 385, pp. 103–104, Jan. 2015.

[246] S. Davies, "Annual Report of the Chief Medical Officer 2016, Generation Genome London: Department of Health," 2017.

[247] G. Wilson, D. A. Aruliah, C. T. Brown, N. P. Chue Hong, M. Davis, R. T. Guy, S. H. D. Haddock, K. D. Huff, I. M. Mitchell, M. D. Plumbley, B. Waugh, E. P. White, and P. Wilson, "Best Practices for Scientific Computing," *PLoS Biology*, vol. 12, p. e1001745, Jan. 2014.

[248] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[249] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, pp. 3111–3119, 2013.

[250] D. Andor, C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins, "Globally Normalized Transition-Based Neural Networks," *CoRR*, vol. abs/1603.06042, 2016.

[251] M. Ford, *The Rise of the Robots: Technology and the Threat of a Jobless Future.* 2015. OCLC: 921985821.

[252] M. D. Workman, *Semantic Web: Implications for Technologies and Business Practices.* 2015. OCLC: 919432525.

Appendices

Appendix A

Appendix A - CRIS Position Paper

BMJ Open

Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource

Gayan Perera,[1] Matthew Broadbent,[2] Felicity Callard,[3] Chin-Kuo Chang,[1] Johnny Downs,[1] Rina Dutta,[1] Andrea Fernandes,[1] Richard D Hayes,[1] Max Henderson,[1] Richard Jackson,[1] Amelia Jewell,[1] Giouliana Kadra,[1] Ryan Little,[2] Megan Pritchard,[1] Hitesh Shetty,[2] Alex Tulloch,[1] Robert Stewart[1]

To cite: Perera G, Broadbent M, Callard F, *et al*. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. *BMJ Open* 2016;**6**:e008721. doi:10.1136/bmjopen-2015-008721

► Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2015-008721).

Received 12 May 2015
Revised 10 November 2015
Accepted 26 November 2015

For numbered affiliations see end of article.

Correspondence to
Professor Robert Stewart;
robert.stewart@kcl.ac.uk

ABSTRACT

Purpose: The South London and Maudsley National Health Service (NHS) Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register and its Clinical Record Interactive Search (CRIS) application were developed in 2008, generating a research repository of real-time, anonymised, structured and open-text data derived from the electronic health record system used by SLaM, a large mental healthcare provider in southeast London. In this paper, we update this register's descriptive data, and describe the substantial expansion and extension of the data resource since its original development.

Participants: Descriptive data were generated from the SLaM BRC Case Register on 31 December 2014. Currently, there are over 250 000 patient records accessed through CRIS.

Findings to date: Since 2008, the most significant developments in the SLaM BRC Case Register have been the introduction of natural language processing to extract structured data from open-text fields, linkages to external sources of data, and the addition of a parallel relational database (Structured Query Language) output. Natural language processing applications to date have brought in new and hitherto inaccessible data on cognitive function, education, social care receipt, smoking, diagnostic statements and pharmacotherapy. In addition, through external data linkages, large volumes of supplementary information have been accessed on mortality, hospital attendances and cancer registrations.

Future plans: Coupled with robust data security and governance structures, electronic health records provide potentially transformative information on mental disorders and outcomes in routine clinical care. The SLaM BRC Case Register continues to grow as a

Strengths and limitations of this study

- Because the Clinical Record Interactive Search (CRIS) model draws directly from the electronic health record, it provides valuable 'real-world' and 'real-time' information on routine mental healthcare, automatically accumulating large volumes of data without any requirement for service reconfiguration or changes at the clinical interface.
- Although electronic health records-based registers remove the requirement for specific 'data collection' in routine clinical care, a major challenge for mental health data in particular is that most information is recorded in text rather than structured fields. Natural language processing offers important opportunities for data enhancement.
- External data linkages are also potentially valuable, but dependent on the nature of the data supplemented—most often providing additional information on exposures and outcomes outside mental health domains and between care episodes rather than on the nature of mental disorders themselves.
- Regardless of the volume of data available, it is important to bear in mind their provenance (ie, highly dependent on what information a clinical staff member records or not); research applications need to be tailored with this in mind.
- A key challenge inherent with all use of healthcare data is data protection, and it is important to develop anonymised data resources in a way that is acceptable to the general public, and to the patients whose personal and often highly sensitive information forms the database. Such challenges incorporate not only a case register's data themselves but also procedures around data linkage where use of identifiers is required.

database, with approximately 20 000 new cases added each year, in addition to extension of follow-up for existing cases. Data linkages and natural language processing present important opportunities to enhance this type of research resource further, achieving both volume and depth of data. However, research projects still need to be carefully tailored, so that they take into account the nature and quality of the source information.

INTRODUCTION

It is nearly 30 years since the publication of Ten Horn *et al*'s[1] comprehensive inventory of the psychiatric case register and its use in research. Seven years ago electronic health record (EHR)-based registers were proposed as a possible 'new generation'.[2] The longitudinal nature of case registers, their size and coverage of defined populations make them an important research asset, providing large numbers of participants and measurement points, as well as the potential for data linkage.[3] Recent years have seen an increase in the use of the psychiatric case register for research purposes, including linkage across diverse health and other population databases, including criminological information resources.[4] There are several unique applications of case registers. Despite the methodological advantages of the randomised controlled trial, observational data remain fundamental to health research, and much of what we know (or assume we know) is derived from observation rather than experimental intervention.[5] Although they can contribute to aetiological research, case registers are particularly suited to the investigation of the course and outcome of a disorder, as well as allowing intervention response to be evaluated in large, naturalistic samples and settings. In smaller scale psychiatric case registers, quality of data can be more regularly checked and the number of variables collected can be higher than in a large database. These registers can include information on the clinical condition of the patients, on psychopharmacological treatments and on duration of contacts.[6] The combination of quality and quantity in data renders small-scale registers of great interest for researchers and policymakers. EHRs in mental healthcare, on the other hand, represent data which are potentially both large and deep—because in theory, these contain every piece of information that has been recorded in a clinical service about a person's presentation, symptoms and relevant background history, as well as interventions received and observed outcomes.[5]

Through technological advances in both the daily updating and validation of registers, large and complex projects can be carried out. Register data are particularly suited to supporting comprehensive longitudinal studies of the course of illness to predict outcomes and naturalistic response to interventions. With EHRs increasingly complementing or replacing handwritten notes in mental health services, large volumes of clinical information are now already contained in an electronic format. This removes the requirement for de novo data collection and entry which presented formidable challenges for earlier registers, albeit processes with a higher potential for quality control. Local EHR-sourced registers are more likely to be limited by migration between geographic catchments, but their strength lies in their ability to cover all types of service within a given area, thereby providing a more comprehensive picture of mental health than is afforded by national registers.

The South London and Maudsley National Health Service (NHS) Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register was set up in 2008 as a novel data resource derived directly from the routine EHRs of a large mental healthcare provider, and its initial development was outlined in 2009.[7] At the time of analysis for that paper (October 2008), the database contained 123 000 cases and information available through the Clinical Record Interactive Search (CRIS) application was primarily restricted to that imposed by the format of the source EHR fields. Since then, the SLaM BRC Case Register has expanded substantially, not only in case numbers (now over 250 000) but also, most importantly, in the scale and depth of derived and externally linked information available. The objective of this paper is to update the description of this case register and, particularly, to outline technical developments which have enhanced the depth of information available, and which we believe have potential generalisability to other comparable clinical data resources.

COHORT DESCRIPTION

The SLaM BRC Case Register and CRIS application

Initial development of the SLaM BRC Case Register has been previously described in detail, as has SLaM as a provider (and see also http://www.slam.nhs.uk).[7] In summary, the data are sourced from EHRs used by SLaM, which provides comprehensive mental health services to a geographic catchment of over 1.2 million residents in four south London boroughs—Croydon, Lambeth, Lewisham and Southwark—as well as some regional/national specialist services. SLaM catchment service provision is currently structured within the following specialty groupings: Addictions; Behavioural and Developmental Psychiatry; Child and Adolescent Mental Health Services; Mental Health of Older Adults and Dementia; Mood, Anxiety and Personality; Psychological Medicine; Psychosis. These are aligned with academic groupings at King's College London, reflecting the university–health service partnership enshrined within King's Health Partners Academic Health Sciences Centre (KHP AHSC; http://www.kingshealthpartners.org; also incorporating two major acute care providers). The groupings also encompass services delivered to all age groups, standard specialties such as Addictions, Eating Disorders and Learning Disabilities, as well as provision within Forensic and General Hospital Liaison

settings. In addition, wider national provision by SLaM at the time of writing includes the following services: adult attention deficit hyperactivity disorder, adult personality disorder, affective disorders, anxiety disorders (residential), autism assessment and behavioural genetics, brain injury (outpatient and inpatient), anxiety disorders and trauma, chronic fatigue, eating disorders (day care, outpatients, inpatients), female hormone clinic, psychosis (inpatient, outpatient and specialist rehabilitation), mother and baby unit, autism, practitioner health, psychological interventions, psychosexual disorders, self-harm (outpatients) and traumatic stress. Finally, some SLaM services provide to a wider geographic catchment (eg, Addiction services to Bexley and Greenwich boroughs) and others are catchment independent (eg, General Hospital Liaison services are provided to the four Acute Trusts within the catchment regardless of individual patients' areas of residence).

Clinical records have been fully electronic (ie, paperless) across all SLaM services since April 2006, using the bespoke Patient Journey System (PJS) which incorporated legacy data from earlier service-specific EHRs. The CRIS application was developed in 2007–2008 and consists of a series of data-processing pipelines which both structure and de-identify PJS fields, rendering effectively anonymised data from the full clinical record available at the researcher interface, with search and database assembly functionality facilitated by a front end designed for non-technical use. The anonymisation process and its effectiveness, including the de-identification of open-text fields and the generation of a pseudonymised identifier (CRIS ID), have been described in detail.[8] The wider patient-led oversight and security model have also been previously described and have not changed significantly since the SLaM BRC Case Register was set up.[7 8] Ethical approval as an anonymised database for secondary analysis was originally granted in 2008, and renewed for a further 5 years in 2013 (Oxford C Research Ethics Committee, reference 08/H0606/71+5). In terms of cohort coverage, all SLaM care is represented on CRIS. An opt-out model is in place for service users, and is advertised in all publicity material and initiatives; to date, only three people have requested this.

The SLaM BRC Case Register conforms to the WHO's formal description of a psychiatric case register—a 'patient-centred longitudinal record of contacts with a defined set of psychiatric services originating from a defined population',[9] although its dynamic nature, updating against source files every 24 h, renders it distinct in some respects. The inclusion of both structured and unstructured (open-text) data in anonymised form, also variously distinguish the SLaM BRC Case Register from other local, regional and national case registries, including those extracted from EHRs such as the disease registries maintained by the US Veteran's Administration.[10 11] Routinely collected data resources such as the Mental Health Minimum Dataset and Hospital Episode Statistics (HES) for England and Wales overlap with SLaM Case Register data but are limited to prespecified structured fields.

Early experience with CRIS and its influence on subsequent design

Developments in the technical architecture underlying CRIS are summarised in the online supplementary appendix and the current model is displayed in figure 1. Studies published to date using CRIS-derived data have generally fallen into two groups. The first have used a combination of open-text and structured data, with open-text data identified using search terms and then manually coded into numeric form for the purpose of analysis. Because of this, sample sizes have been limited to no more than several hundred. However, productive examples include one of the largest case series assembled of people with neuroleptic malignant syndrome, in order to evaluate the range of diagnostic criteria,[12] and associations with antipsychotic exposure,[13] as well as a study of factors associated with khat use in a comprehensive sample of Somali mental health service users.[14] The second group of studies have used only structured data or have made very limited use of open-text data. These have typically analysed sample sizes of several thousand or more. Examples include studies of residential mobility and of homelessness among inpatients on mental health wards, and a series of investigations of mortality associated with mental disorder, described later.[15 16]

Important experiential learning occurred during the initial stages of CRIS use. First, we found that it was sometimes desirable to select and combine data from records in ways that were unsupported by the original CRIS interface (eg, because of complex temporal relationships required between fields). Second, it became clear that while being able to identify and retrieve open-text records according to the presence of prespecified search terms did achieve helpful economy of effort, it did not remove the work needed to generate quantitative data from open text. Indeed, for those projects dependent on the use of open text, the manual coding process placed important limitations on sample size and study duration. Finally, researchers began to develop ideas that required data in addition to those stored in the source EHR, such as data from primary care, acute care and outcomes such as mortality. In the succeeding sections, we set out how the SLaM BRC Case Register has evolved to respond to these challenges.

Handling open text

As outlined above, a priority for development has been to develop more efficient ways of using open-text data in the SLaM BRC Case Register. Early case register data collection included manually reading the de-identified text fields returned by CRIS, such as routine case notes, correspondence and medication notes. For example, one of the recent publications involved manually reading of 2860 records on CRIS of patients receiving

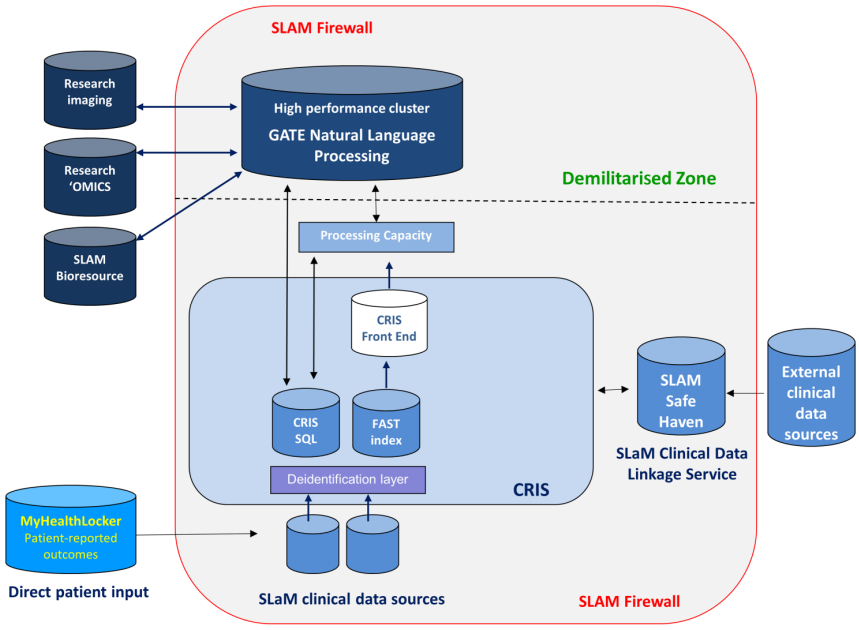


Figure 1 Diagram/map of CRIS technical architecture including natural language processing and data linkage. CRIS, Clinical Record Interactive Search; GATE, General Architecture for Text Engineering; SLaM, South London and Maudsley.

acetylcholinesterase inhibitors in order to record their Mini-Mental State Examination (MMSE) scores and respective dates, and other medication prescribed.[17] Through this process over 11 000 MMSE scores were ascertained; however, there were significant demands in terms of time and resources and the exercise was only possible as the focus of a PhD studentship. Beyond the efficiencies in manual coding gained by extracting only those records required for coding, through keyword searches and postsearch processing, further gains may be made by displaying text fields in ways that make text of interest easier to see, and by displaying data that are required to be reviewed together in close proximity, and away from other data. For example, in studies of homelessness and residential mobility among inpatients, 4485 admissions were selected according to defined criteria, and free-text records corresponding to these admissions were selected if they contained the terms 'homeless', 'NFA' or 'no fixed abode'.[15 16] The aim was to check structured data on homelessness against free-text data, and if necessary, to supplement the former. SAS was used to insert 'tags' that change font colour (red) and weight (bold) for the target words when the data are displayed in Excel, allowing around 2000 free-text progress notes to be coded as homeless/not homeless in less than a day. A SAS Enterprise Guide project developed in collaboration with Amadeus Software Ltd allows CRIS users to do this via a graphical user interface.

A more ambitious approach has been followed for an ongoing project to capture incident cases of psychosis, supported by another Enterprise Guide project developed in collaboration with Amadeus Software Ltd. First, a structured query language (SQL) query retrieves a selection of data for individuals not already present on a cumulative database of first-episode psychotic patients and not already diagnosed as having a psychotic disorder, and whose recent free-text entries contain particular words of interest such as 'delusion' or 'hallucination'. Second, these data are imported into SAS and then automatically outputted in a format suitable for manual coding. This involves splitting data into a multiworksheet Excel workbook, such that each worksheet (tab) contains only data relating to a single person (in the case of our proposed project, each worksheet would similarly pertain to a single episode of care). Targeted words are displayed in colour and in bold.

In contrast to the facilitated, but still manual approaches described above, natural language processing (NLP) techniques have been evaluated and applied for extracting knowledge from unstructured text data. For our purposes, the key NLP technique has been information extraction (IE) where unstructured text is converted into structured tables.[18] Such methods promise massive reductions in the time resource required by researchers to unlock information held in clinical notes that in turn may be connected to other parts of the structured record. It was therefore decided, early in the postdevelopment phase, to implement a text-mining capability in CRIS. This was to be generic, in that information to be extracted could not necessarily be foreseen in advance of the design of individual research studies. General Architecture for Text Engineering (GATE) was chosen as the core NLP infrastructure for CRIS.[19 20] GATE is a widely used suite of open source software for text engineering that includes a workbench for developing applications, tools for distributing those applications on different computer hardware architectures, a quality assurance suite and facilities

for manual preparation of example data.[19–21] GATE's origins are in clinical IE and it has been widely applied in this context.[22 23] GATE includes a flexible architecture for IE and text mining, a large set of pluggable text processing components, and graphical tools for organising those components into new applications. The GATE suite also includes tools for text-mining workflow, distributed processing and visualisation. A variety of text processing tools and document formats may be plugged into this architecture, with individual tools being chained together into processing 'pipelines', and documents processed in series through these pipelines.

Two distinct shallow language processing methodologies have been adopted for CRIS development, in collaboration with University of Sheffield Department of Computer Science. The first may be described as rule-based pattern matching of key concepts. Sentences are first processed to find and create annotations based on simple surface linguistic information (such as words, sentences, etc). This step is then followed by the process of finding concept-specific keywords, which are used to recognise likely sentences of importance to the IE task. For example, in an application to determine the smoking status of a patient implied by texts, such a dictionary, might list the terms of common tobacco products and activities—'cigarette', 'smoker', etc. Finally, a set of patterns specific to the text-mining task are run over the previously generated annotations in order to create a final annotation containing all of the information required in a readily extractable format. The challenge of the pattern matching approach is that it is knowledge intensive. A successful series of patterns need to be developed in relation to a specific IE task (eg, to extract medications, educational level or particular test results). They have to be built manually by GATE users with language engineering skills, using definitions agreed with clinicians and epidemiologists. A sample of the output from an initial prototype application is then corrected by a clinician or epidemiologist, which in turn is used to stimulate discussion about requirements and to provide a basis for multiple iterations of development until performance requirements are met. An advantage of this IE approach is that it also allows researchers to combine information available from open text and structured fields available in CRIS, through SQL, thus combining multiple sources of information. At the postprocessing stage, we can further apply specific filtering criteria to data extraction, such as frequency and length of prescribing and number of concomitant drugs, thus identifying more complex patterns in the text, such as antipsychotic medication profiles (ie, antipsychotic polypharmacy).[24]

Because of the lengthy development cycles of building shallow parsing algorithms, a second IE methodology has also been evaluated. Here, support vector machines (SVMs) are used to rapidly achieve respectable results for certain types of IE problem. A SVM is a machine-learning technique where the intention is to represent instances of text as vectors in high dimensional space. With a training set of instances labelled as indicative of a desired class, the SVM implementation in GATE generates a hyperplane which can in turn be used to classify unseen instances pertaining to the described class in the training set. In practice, this primarily uses a technique known as 'bag of words', where the occurrence of single words within a sentence is the principal currency used to distinguish the various classes. The first part of the model construction requires an expert (eg, clinician) to review a set of documents and label sentences which are relevant to the concept in question, in much the same way that they might signal to a language engineer the relevance of a given sentence for a pattern-based approach. The combination of labelled and unlabelled sentences forms the training data, from which the SVM learns the classification function. This model is then applied to unseen data, and the model quality assessed by human review. If required, further training data can be supplied, which may involve an active learning-inspired approach. A limitation with SVMs applied in CRIS has been that they have limited suitability for complex data extraction problems; however, in scenarios where the assertion to be extracted is simple and tend to be restricted to a concise set of clinical language, performance has been found to be very good and IE applications with immediate utility can be rapidly developed.[25] The TextHunter program was designed specifically to aid the process of clinical text annotation in CRIS, providing an easy-to-use interface for annotators with a focus on the sentence containing the word(s) of interest and immediately proximal text and functionality for rapid coding into discrete groups, typically comprising the following: (1) positive (ie, implying that the construct is present); (2) negative (ie, a statement indicating that the construct is absent); and (3) irrelevant text.[26] Additional TextHunter functionality includes platforms for interannotator agreement testing, and the creation of gold standard and test annotation sets.

Whether rules-based or machine-learning approaches are used, separate training and test data sets are constructed. Standard metrics for evaluating IE application performance in the test data sets, at the level of the individual text annotation, comprise **precision** (equivalent to positive predictive value; the proportion of IE application 'hits' which are found to identify the genuine construct) and **recall** (equivalent to sensitivity; the proportions of instances of the genuine construct which are identified by the application). Employing text mining within the CRIS data set has involved a trade-off between the two. However, the longitudinal nature of EHR data means that there are generally multiple opportunities for an NLP application to capture a piece of information; therefore, suboptimal recall can be compensated for and the focus has been on maximising precision. For the purpose of precision and recall testing, there are two reportable outcomes. The first is 'annotation level', which is carried out across randomly selected

documents and is an indicator of the base level of performance of the application. This figure is useful for developmental purposes, or, in the case of simple concepts that do not require postprocessing, for estimating the final performance of the algorithm. The second type of precision and recall are 'currency level', measuring performance after postprocessing.

The SLaM Clinical Data Linkage Service

SLaM comprises one part of the KHP AHSC (established with King's College London, Guy's and St Thomas' and King's College Hospitals NHS Foundation Trusts) and received National Institute of Health Research (NIHR) funding to set up a service to meet the growing demand from SLaM and KHP researchers whose projects require linked data extracts. SLaM consequently established the Clinical Data Linkage Service (CDLS) as a trusted third party safe haven set up to enable safe and secure data processing services (linkage, and/or storage, and/or extraction) on distinct data sets for secondary research use. The two main methods of linkage have involved either (1) CDLS performing a secure linkage using deterministic or probabilistic matching if/as required or (2) CDLS supporting another trusted third party service to perform the linkage outside of the SLaM electronic firewall followed by CDLS receiving the linked data afterwards (eg, CRIS-HES linkage). Linked data are stored by CDLS in accordance with the SLaM ICT Security Policy and a set of standards contained in a CDLS Memorandum of Understanding completed by the data controllers providing data to individual projects, prior to undertaking any data processing for the project. Linked data are stored on a CDLS server within the SLaM firewall. To date, linkages have been successfully carried out between CRIS and a number of databases, described below.

Primary care (Lambeth DataNet)

Lambeth DataNet (LDN) has been used for several research studies.[27 28] Using the services of a contracted partner, Quality Medical Solutions (QMS) until April 2014, data are extracted and pseudonymised from the general practitioner (GP) practices in question. In terms of the mechanism of linkage, QMS scramble the patient identifiable information (NHS number) within the complete LDN data set and send the algorithm to the CDLS using an official encrypted NHS data transfer method to allow linked data files to be generated within CDLS. All identifying data other than CRIS and LDN pseudonyms are then removed. On final approval, SLaM BRC researchers will submit their data extract request to CDLS, either using CRIS to identify a discrete list of client pseudonyms for their project cohort to be linked with CRIS and LDN data (this pseudonym is not returned to the researcher), or submitting a detailed description of the cohort under investigation for CDLS to assemble the corresponding linked data. Once the linkage is complete, the LDN ID pseudonym is destroyed and an anonym (project-specific ID) is used thus creating a project-specific, fully anonymised data set for analysis. LDN currently extracts data from all GP practices in Lambeth—that is, around a quarter of the geographic catchment served by SLaM.

Department for Education National Pupil Database

The Education (Individual Pupil Information; Prescribed Persons; England) Regulations 2009 as amended by The Education (Individual Pupil Information; Prescribed Persons; England; Amendment) Regulations 2013 enable the Department for Education (DfE) to share individual pupil information from the National Pupil Database (NPD) with named bodies and persons who, for the purpose of promoting the education or well-being of children in England, are conducting research or analysis, producing statistics, or providing information, advice or guidance. Access is subject to requesters complying with terms and conditions imposed under contractual arrangements and a rigorous approvals process. The DfE Data Management Advisory Panel approved the DfE Data and Statistics division linkage service to undertake the linking of IDs between CRIS and the NPD. In terms of the data linkage mechanism, SLaM CDLS will first identify all children under 17 on the CRIS database, comprising approximately 35 000 cases who have attended SLaM Children and Adolescent Mental Health Services between 1 January 2008 and 31 December 2013. Identifiers will then be sent via secure file transfer to the DfE Data and Statistics Department who will match these against the NPD identifiers cohort (approximately 15 million records), generating a pupil-specific, non-identifiable NPD ID variable across the whole data set, and adding the CRIS ID to this table for cases only, stripping the resultant table of all identifiers other than the anonymised NPD ID and the pseudonymised CRIS ID, and transferring the data set back to SLaM CDLS using secure file transfer. Researchers on approved projects will compile clinical data from CRIS for approved analyses and send to CDLS for linking. CDLS will then fully anonymise resultant tables by replacing the CRIS ID for cases throughout with a project-specific CDLS ID, and the link between the CRIS ID and CDLS ID will be permanently destroyed prior to sending linked tables to researchers for analysis.

Hospital Episode Statistics

HES data are compiled from all NHS Trusts in England (both acute and mental health services), including statistical abstracts of records of all inpatient episodes, as well as outpatient and emergency care. For this linkage, CRIS identifiers are compiled by CDLS, and transferred to the Health and Social Care Information Centre (HSCIC) using an NHS-approved secure file transfer protocol. HSCIC then adds the CRIS ID to all HES records that match CRIS records and extracts all other HES records for patients within the four catchment

boroughs served by SLaM (the control group). HSCIC destroys patient identifiers leaving only the CRIS ID and HES extract ID. As with other linked data sets, the CRIS-HES data are transferred back to CDLS to be held and provided to researchers in a fully anonymised format.

Mortality

Office for National Statistics (ONS) mortality data are additionally requested via the HSCIC. CDLS send identifiers (CRIS ID, first name, last name, date of birth, gender, postcode and NHS number) to HSCIC, who return ONS mortality data to CDLS via the same secure file transfer protocol as that used for the HES linkage. While ONS mortality data include details of information recorded on the death certificate, date of death is available on a wider CRIS sample through data held by SLaM, in common with most mental health NHS Trusts through standard linkage of all NHS numbers to the national spine.

Cancer

In an initial piece of work, a data linkage was set up between CRIS and Thames Cancer Register by the UK Government Department of Health Research Capability Programme, findings from which have been previously reported and which generated an irreversibly anonymised linked data set.[29] This data resource is currently being expanded to bring together updated local data from the National Cancer Registration Service (NCRS) held by Public Health England's London Knowledge and Intelligence Team, linking this with CRIS and incorporating additional HES and mortality data provided by HSCIC and ONS.

Procedures and resources

Results from all these linkages are stored within the CDLS safe haven, and CDLS plays a key role in wider governance, supplementing the role of CRIS-specific oversight and data security previously described.[7 8] While set up to support research at the SLaM BRC, as an independent trusted third party service CDLS sits outside the BRC and is managed by a dedicated team within the SLaM Information and Communications Technology department, reporting directly to the SLaM Director of ICT Strategy and ultimately accountable to the SLaM Trust Board. Important features of CDLS work are the secure handling and storage of identifier fields required for data linkage. Section 251 (s.251) of the NHS Act 2006 allows the common law duty of confidentiality to be set aside in specific circumstances where anonymised information is not sufficient and where patient consent is not practicable. S.251 approval has been granted to SLaM for all the above linkages, which allow data to be available in an identifiable format to a small number of data processing staff in accordance with data sharing contracts. Activity for projects using linked data sets held by CDLS is audited by the CDLS Safe Haven Officer, helping to ensure that the user's project requirements (eg, clinical research, surveillance, service improvement or audit) are met, and projects progress within the agreed policy and practice framework. The CDLS communications plan has a patient-facing aspect in raising awareness of the projects facilitated by the CDLS. Service user involvement is ensured in the decision-making process of approving projects working with linked data held by CDLS, and the patient-chaired CRIS Oversight Committee reviews and approves all projects using CRIS-linked data. Separate committees with the same terms of reference have been set up to provide governance for the LDN and NPD linkages, in order to accommodate representation from respective agencies providing these data.

Four distinct services are thus offered by the CDLS. First, CDLS provides advice on permissions, approvals and contracts. These include consideration of academic, technical, legal and ethical requirements. The SLaM 'Caldicott Guardian' is responsible for any use of patient identifiable information and their approval is also a prerequisite. Second, CDLS facilitates data linkages either within the CDLS safe haven or via a third party, coordinating the secure transfer of data. Third, CDLS is responsible for the secure storage of linked data in accordance with predefined information governance and security standards. Fourth, CDLS as the custodian for the linked data prepares and extracts bespoke and prespecified databases for approved CRIS projects and provides these to researchers. Therefore, there is no direct access by researchers to the full linked data files, enhancing data protection and confidentiality.

Cohort characteristics

Initial descriptive data were assembled on the catchment area for SLaM (Croydon, Lambeth, Lewisham and Southwark) using publicly available sociodemographic information from ONS census data.[30] Analyses of CRIS data used 31 December 2014 as a census date for descriptive statistics including sociodemographic and diagnostic profiles. 'Active' patients on this date were defined as those who had been referred to and accepted by SLaM and had not been discharged by 31 December 2014. 'Inactive' patients had a recorded activity date on or before 31 December 2014 and excluded referrals categorised as 'rejected' or 'waiting'. On 31 December 2014, 223 224 patient records were available on CRIS, of which 31 961 described 'active' patients and 191 263 'inactive'. The remaining 21 882 records described referrals, which were either solely characterised as 'rejected' or 'waiting', and in which no team episode (for outpatients) or ward stay (for inpatients) was indicated. Descriptive data were further provided for key linked data sets at that time. In this respect, the most recent mortality date recorded in the linked ONS mortality data set was 16 December 2013; cancer registry data were linked up to 31 December 2008; HES data were available to 31 March 2013. For analyses of linked HES

data, contacts with mental health services were excluded.

Descriptive data from the UK Census for the catchment populations served by SLaM are summarised in table 1 and contextualised with the same information for London as a whole and for England. There are slight differences in population structure between the four boroughs served, with Croydon having higher proportions of young children and older residents compared with London and the other three boroughs. Highest proportions in the young adult (20–39 year) age range were living in Lambeth and Southwark. As a whole, the SLaM catchment has a slightly higher predominance of working adults in the 20–59-year range compared with London, and shares with London lower proportions in older age ranges compared with England. The SLaM catchment has substantially higher proportions of residents from minority ethnic groups and/or born outside UK compared with England, whereas compared with London as a whole, there are higher proportions from black minority groups and lower proportions from Asian groups. In common with London as a whole, proportions are higher in both highest and lowest socioeconomic groups compared with England; proportions in unemployment are higher, but so are proportions with higher levels of education. Of the catchment boroughs, Lambeth, Southwark and Lewisham have higher levels of both in-migration and out-migration compared with Croydon. Based on the ratios between summed borough statistics and those for the catchment overall, 76.9% of inflow migration and 78.5% of outflow migration was from/to areas outside the catchment, rather than between catchment boroughs.

Geographic characteristics are summarised in figures 2–4. Figure 2A visually contextualises deprivation levels in SLaM compared with other areas of London, and figure 2B summarises the most recently recorded residence of active SLaM patients. In the latter, most active SLaM patients were identified as residing within its geographic catchment, although appreciable numbers were drawn from a wider geography. Within the SLaM catchment, higher numbers of active patients were generally found in areas of higher deprivation, although several anomalous areas can be seen—for example, those with high deprivation and relatively low numbers of active patients (figure 3A, B). Figure 4 illustrates the most recent recorded residence of non-active patients in London (figure 4A) and specifically in SLaM's catchment (figure 4B). Outside SLaM's catchment, relatively high numbers of inactive patients were recorded as residing in neighbouring local authorities in South East London including Bexley, Greenwich and Bromley.

Descriptive data are summarised in table 2 for all people who were represented on the SLaM BRC Case Register on 31 December 2014. Higher proportions of active patients were 80 years and older and in the 40–59 group compared with proportions in the four catchments. Compared with the catchment area characteristics described in table 1, active SLaM patients had a slightly higher male predominance, and there were higher proportions self-assigning as white, mixed or other ethnicity. Around 70% were single. Employment status data were available on less than 25% of the active sample, but of this group around 66% were unemployed. Of active SLaM patients on the census date, 6574 (20.9%) were either residing in boroughs outside London or living in London but outside SLaM's four catchment boroughs. Of these, 3385 (51.5%) were in contact with SLaM services that provided for other boroughs, 1941 (29.5%) were using one or more of SLaM's national services, 341 (5.2%) were in contact with General Hospital Liaison services covering one of the four Acute Trusts within the SLaM catchment, and 907 (13.8%) were previous catchment residents currently living outside the catchment (193 of whose addresses were recorded as temporary).

On the 31 December 2014 census date, there were nearly 32 000 active cases receiving care from SLaM services, with the largest numbers receiving care from Psychosis or Child and Adolescent Mental Health Services (table 3). A further 190 000 plus patients on the SLaM BRC Case Register were inactive to SLaM, nearly one-third of whom received care from Psychological Medicine services (which includes General Hospital Liaison services). Table 4 provides an additional description of overlap between services for active and inactive patients, with over 1000 active patients in contact with two or more specialties concurrently and over 15 000 inactive patients having received care from two or more specialties. Ever-recorded primary diagnoses are summarised in table 5. Of active patients, the most common mental disorder diagnoses ever recorded were schizophrenia (21.2%) and mood (19.0%) disorders, followed by organic (11.0%), substance use (11.7%) and neurotic (13.0%) disorders, and disorders of childhood and adolescence (11.3%). Sizes of data linkage samples are described in tables 6–8. Nearly 85% of CRIS patients had records in HES (excluding mental health service data) and nearly 2% of CRIS patients had data linked to those from the cancer registry within the years of data availability (table 6). Distributions of underlying cause of death are summarised in table 9 for the linked sample with this information, and primary cancer diagnoses are similarly described in table 10.

Performance of NLP applications

Performances of IE applications to date are summarised for CRIS as a whole, supplementary to more detailed publications on some of these.[31 32 33 34] The first NLP IE application to be developed was for the MMSE, a commonly used 0–30-point assessment of global cognitive function. The objective of the application was to ascertain both the numerator and denominator scores (because denominator scores of less than 30 are used where some items cannot be attempted

Table 1 Descriptive statistics, derived from the 2011 UK Census, for the four London boroughs served by SLaM, compared with statistics for London and England as a whole

	SLaM catchment					Comparison statistics	
	Lambeth	Croydon	Lewisham	Southwark	Combined	London	England
Total population*	310 200	368 900	281 600	293 500	1 254 200	8 308 400	53 493 700
Age (%)							
<20	21.7	26.9	25.4	23.0	24.4	24.5	24.0
20–39	44.2	29.3	36.3	41.7	37.5	35.8	27.0
40–59	23.4	26.9	25.3	24.4	25.1	24.5	26.7
60–79	8.6	13.5	10.3	8.8	10.4	12.1	17.7
≥80	2.1	3.4	2.7	2.1	2.6	3.1	4.6
Gender (%)							
Male	49.8	48.5	48.9	49.5	49.1	49.3	49.2
Female	50.2	51.5	51.1	50.5	50.9	50.7	50.8
Education† (%)							
No qualifications	14.2	17.6	17.7	16.3	16.5	17.6	22.5
Highest level of qualification; level 1 qualifications	8.5	13.8	11.1	9.4	10.9	10.7	13.3
Highest level of qualification; level 2 qualifications	9.8	15.2	12.5	10.2	12.1	11.8	15.2
Highest level of qualification; apprenticeship	1.1	2.1	1.4	1.2	1.5	1.6	3.5
Highest level of qualification; level 3 qualifications	9.7	11.4	10.8	10.5	10.6	10.5	12.4
Highest level of qualification; level 4 qualifications and above	46.6	31.8	38	43.1	39.5	37.7	27.4
Highest level of qualification; other qualifications	10.1	8.1	8.5	9.3	9.0	10.1	5.7
Self-assigned ethnicity (%)							
White	57.1	55.2	53.5	54.3	55.1	59.8	85.5
Mixed	7.6	6.4	7.4	6.2	6.9	5.1	2.2
Asian or Asian British	6.8	16.4	9.3	9.5	10.8	18.4	7.7
Black or Black British	25.9	20.2	27.2	26.8	24.7	13.3	3.4
Other	2.6	1.8	2.6	3.2	2.5	3.4	1.2
Socioeconomic classification (%)‡							
Higher managerial, administrative and professional occupations	16.2	14.1	13.1	15.8	14.8	15.8	13.8
Lower managerial, administrative and professional occupations	27.3	24.8	25.7	24.8	25.6	24.7	22.8
Intermediate occupations	10.6	13.7	12.1	10.3	11.8	10.9	10.5
Small employers and own account workers	9.7	12.9	10.9	8.8	10.7	12.9	12.8
Lower supervisory and technical occupations	5.9	6.7	6.8	6.6	6.5	6.5	8.8
Semiroutine occupations	10.3	12	12.6	12	11.7	10.9	13
Routine occupations	9.7	8.3	8.7	9.9	9.1	8.8	12.1
Never worked and long-term unemployed	6.9	5.3	6.4	7	6.3	6.5	4.2
Full-time students	3.4	2.2	3.7	4.8	3.4	3	2
Percentage of people born in UK	61.1	70.4	66.4	63.2	65.5	85.8	94.1
Estimated migration (thousands per year) for the 1 year period ending June 2014§							
Inflow	29.07	19.19	21.2	25.25	72.81	196.6	526
Outflow	31.78	19.81	22.36	27.53	79.71	251.6	314
Balance	−2.71	−0.62	−1.16	−2.28	−6.90	−55	+212

*Resident population estimates by broad age band, mid-2013, using ONS 2011 census.
†All usual residents aged over 16 on the census date 27 March 2011.
‡Based on HRP: an individual person within a household to act as a reference point and charactering whole household according to characteristics of the chosen reference person.
§Data source: http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-326817 accessed on the 5 November 2015. SLaM catchment and London statistics calculated for the 1 year period ending June 2013 (and the overall catchment statistic does not include within-catchment migration); England figures represent rolling annual data for year ending June 2014.
HRP, household reference person; ONS, Office for National Statistics; SLaM, South London and Maudsley.

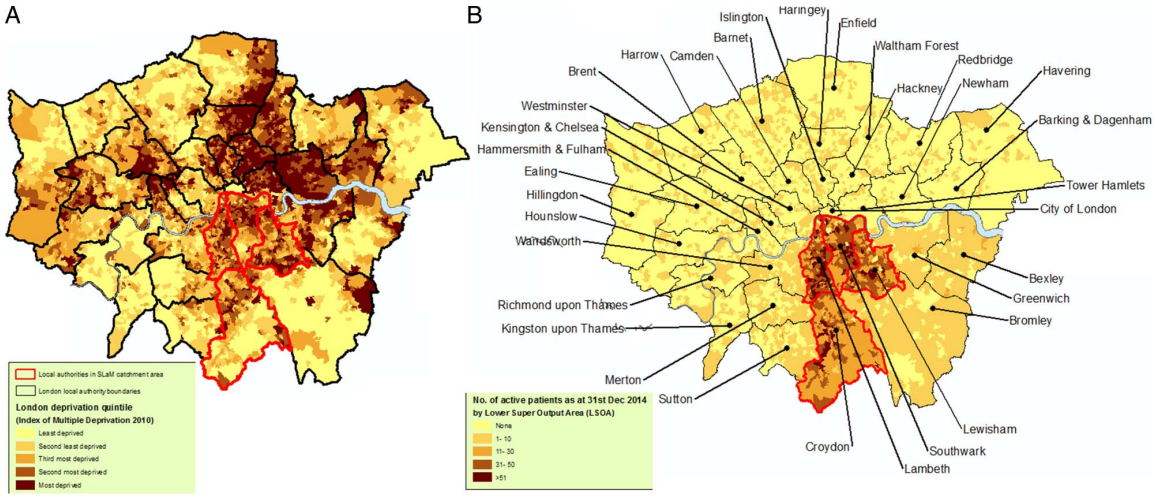


Figure 2 Maps contextualising deprivation levels in the South London and Maudsley (SLaM) catchment compared with London as a whole, and illustrating the distribution of recorded residences for active patients (on 31 December 2014) within London.

because of, eg, sensory impairment), as well as the date implied for the assessment (because clinical text fields commonly refer to previous as well as current scores). Further rules for application postprocessing were that only MMSE scores with denominators over 25 were included (because scores below that level imply substantial missing data and a scale that was probably incompletely administered), and scores were excluded if two different numerators were assigned to the same date.[34] The application for educational attainment sought to ascertain the numeric value associated with text commenting on school leaving age, whether the age itself or the year, and the application for 'living alone' simply sought to identify that phrase or equivalents applied to the patient. In developing the smoking application, authors extracted information from open-text fields, classifying patients as either 'currently smoking', 'past smoker' or 'has never smoked', with smoking of substances other than tobacco (eg, marijuana/cannabis and cocaine) specifically excluded.[31] The methodology used an iterative process of manual 'gold standard' annotation of free-text documents, followed by comparison with the results generated by the application at each development stage, with analysis of this comparison feeding further development of the rules. The application for 'diagnosis' sought simply to extract any text strings associated with a diagnosis statement in order to supplement the

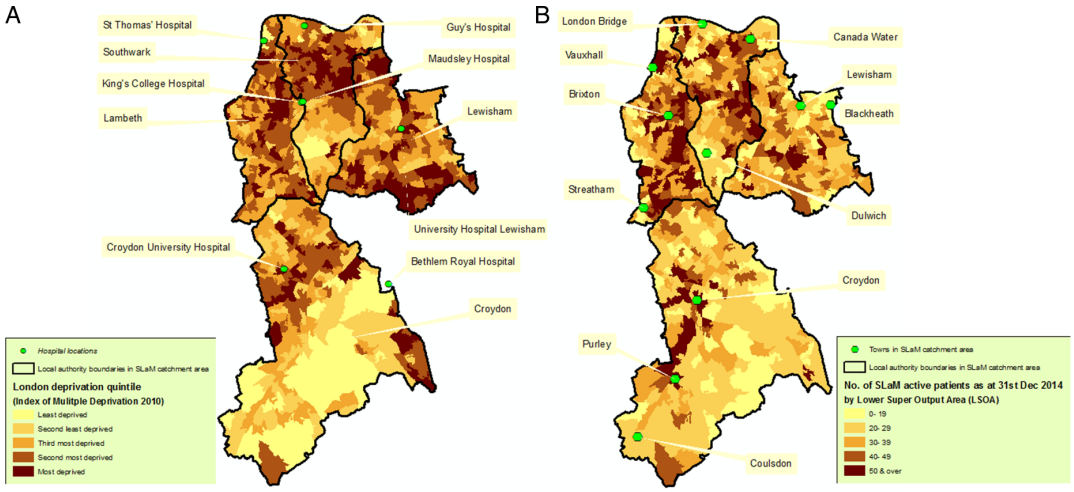


Figure 3 (A, B) Maps showing distribution of deprivation levels in the four catchment boroughs served by South London and Maudsley (SLaM), the key hospital sites and the number of active patients (on 31 December 2014) across the same geography.

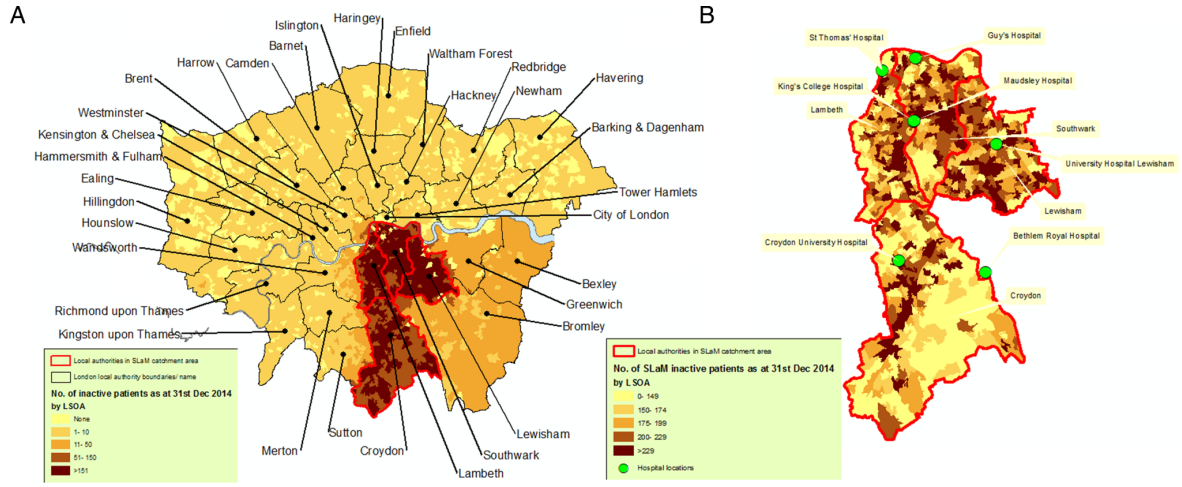


Figure 4 Maps illustrating the distribution of recorded residences for inactive patients (on 31 December 2014) within London and SLaM catchment area. LSOA, lower super output area; SLaM, South London and Maudsley.

existing structured (International Classification of Diseases (ICD)-10) fields. Its performance was evaluated formally in a random sample of 75 documents for 'vascular dementia',[33] but is recommended for individual further evaluation in other conditions. The application for ascertaining pharmacotherapy was developed using a gazetteer of generic and commercial names for all medications in UK use in order to ascertain instances where the patient was reported as receiving these, with supplementary rules for ascertaining recorded dose, frequency/timing and starting/stopping statements. Its precision was first tested for clozapine receipt against a manual search of 279 documents, and recall was ascertained on a random set of 200 documents containing the word clozapine and scrutinised to ascertain an actual prescription.[32] Finally, the validity of this application was recently further evaluated for six antipsychotic agents (amisulpiride, flupentixol, haloperidol, olanzapine, risperidone, zuclopenthixol) on instance level (ie, specific mentions in the text at individual points in time). To estimate precision and recall, the authors examined a subset of 20 patients for each medication, totalling 120 patients (the instances of antipsychotic prescribing varied from 328 to 1150 instances by antipsychotic agent) by running the NLP application over the set of unseen documents and comparing the results to the manual coding of the same data set.[24] For all evaluations, an F-statistic was additionally calculated, representing the harmonic mean of precision and recall, and defined as: $F=2\times(\text{precision}\times\text{recall}/(\text{precision}+\text{recall}))$. As with the diagnosis application, further bespoke validation of the pharmacotherapy application is recommended for new medications or classes. Performance data are summarised for NLP IE applications in table 11, and table 12 describes the resulting additional structured data points generated across CRIS using these applications.

Findings to DATE

The SLaM BRC Case Register has been used for a wide range of research projects to date, as well as for key service evaluation and audit projects, and over 50 publications have arisen. Large-scale outcome studies supported by CRIS data have included those of residential mobility and of homelessness among inpatients on mental health wards.[15 16] Evaluations of service interventions and other quality markers were also studied,[35 36] and investigations are increasingly focusing on early symptoms and treatment pathways in psychosis.[37 38] Keyword search functionality recently supported a large historic cohort study of service use and abuse experiences of trafficked people in contact with secondary mental health services.[39]

A particularly prominent theme has been the investigation of mortality and physical health outcomes in people with mental disorders. Initial reports highlighted the raised mortality and lower life expectancy of people in the most common disorder groups.[40–43] More studies were carried out to attempt to profile those most at risk, which have indicated that disability and environmental circumstances appear to be more important than symptoms.[44 45] This was supported by a study showing that, in those who received specific structured risk assessments, clinician-perceived risk of self-neglect was a strong and independent predictor of mortality, whereas clinician-perceived risks of suicide and/or violence were not predictive.[46] In terms of mortality predictors in specific patient groups, the impact of psychiatric comorbidity and psychological health on all-cause and cause-specific mortality in opioid use disorder has been evaluated, highlighting the importance of personality disorder and comorbid alcohol use disorder.[43] Similarly, the importance of alcohol and drug use, physical illness, and functional impairment as predictors of mortality in individuals with personality disorder has been

Table 2 Characteristics of patients represented on the South London and Maudsley National Health Service (NHS) Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register records (census date: 31 December 2014)

Characteristic	Active patients (%) N=31 961	Inactive patients* (%) N=191 263
Current age (years)		
<20	6265 (19.6)	23 740 (12.4)
20–39	9464 (29.6)	65 493 (34.2)
40–59	10 101 (31.6)	59 336 (31.0)
60–79	4017 (12.6)	23 924 (12.5)
≥80	2114 (6.6)	18 770 (9.8)
Year of birth		
On or after 1994	6785 (21.2)	27 214 (14.2)
1993–1973	10 032 (31.4)	68 722 (35.9)
1973–1954	9337 (29.2)	53 317 (27.9)
1953–1934	3913 (12.2)	22 065 (11.5)
On or after 1933	1894 (5.9)	19 945 (10.4)
Gender		
Male	16 780 (52.5)	93 902 (49.1)
Female	15 160 (47.5)	97 327 (50.9)
Self-assigned ethnicity (full breakdown)†		
British	14 833 (50.5)	83 425 (55.6)
Irish	614 (2.1)	3819 (2.5)
Any other white background	2196 (7.5)	13 072 (8.7)
Mixed: white and black	770 (2.6)	2899 (1.9)
Mixed: white and Asian	104 (0.4)	421 (0.3)
Mixed: any other mixed background	277 (0.9)	961 (0.6)
Indian	413 (1.4)	2072 (1.4)
Pakistani	211 (0.7)	958 (0.6)
Bangladeshi	115 (0.4)	631 (0.4)
Any other Asian background	596 (2.0)	3105 (2.1)
Caribbean	2192 (7.5)	7654 (5.1)
African	2156 (7.3)	9178 (6.1)
Any other black background	2923 (10)	10 628 (7.1)
Chinese	107 (0.4)	593 (0.4)
Any other ethnic group	1865 (6.3)	10 715 (7.1)
Ethnicity not known or not stated	2589 (8.8)	41 132 (21.5)
Self-assigned ethnicity (amalgamated)†		
British, Irish or any other white ethnic groups	17 643 (60.1)	100 316 (52.4)
Mixed	1151 (3.9)	4281 (2.2)
Indian, Pakistani, Bangladeshi or 'other Asian'	1335 (4.5)	6766 (3.5)
Caribbean, African or any 'other black'	7271 (24.8)	27 460 (14.4)
Other	1972 (6.7)	11 308 (5.9)
Area of most recently recorded residence‡		
Croydon	6127 (19.5)	36 996 (20.4)
Lambeth	7043 (22.4)	33 471 (18.5)
Lewisham	5610 (17.8)	35 206 (19.4)
Southwark	6120 (19.4)	32 961 (18.2)
Other London boroughs	3179 (10.1)	27 012 (14.9)
Outside London	3395 (10.8)	15 649 (8.6)
Unknown	487 (1.5)	9968 (5.2)
Most recent employment status		
Paid employment	439 (6)	5118 (13.4)
Part-time employment	114 (1.6)	581 (1.5)
Self-employed	31 (0.4)	408 (1.1)
Volunteer	67 (0.9)	95 (0.2)
Government training scheme	<10 (0.1)	24 (0.1)
Full-time student	204 (2.8)	1623 (4.2)
Full-time student—school age	930 (12.7)	7725 (20.2)
Retired	504 (6.9)	6790 (17.8)
Registered disabled	71 (1.0)	352 (0.9)

Continued

Table 2 Continued

Characteristic	Active patients (%) N=31 961	Inactive patients* (%) N=191 263
Unemployed	4827 (66.1)	14 949 (39.1)
Other	115 (1.6)	534 (1.4)
Employment status not known	24 654 (77.1)	153 064 (80.0)
Most recent marital status		
Married	1329 (4.7)	13 701 (10.4)
Married/civil partner	3111 (11)	11 027 (8.3)
Cohabiting	556 (2.0)	2532 (1.9)
Divorced	622 (2.2)	3920 (3.0)
Divorced/civil partnership dissolved	633 (2.2)	2293 (1.7)
Separated	853 (3.0)	5303 (4.0)
Widowed	320 (1.1)	5985 (4.5)
Widowed/surviving civil partner	1046 (3.7)	4280 (3.2)
Single	19 763 (70)	83 319 (62.9)
Marital status not known or not disclosed	3728 (11.7)	58 903 (30.8)

*Inactive: those not currently receiving treatment and who have been discharged from all services.
†Excluding those not stated or none: active=2589/31961, inactive=41 132.
‡As at 31 December 2014.

demonstrated, a group with and life expectancies at birth reduced by 17–19 years compared with the general population in England and Wales.[47 48] Mortality outcomes have been further evaluated in studies of cognitive impairment and delirium in older adults.[34 49]

Studies of pharmacotherapy profiles have continued investigations into mortality as an outcome, most notably in a report identifying a marked reduction in people using clozapine, not explained by a range of potential confounders including service use.[32] Another study found that atypical antipsychotic agents were not associated with higher mortality in people with vascular dementia.[33] Further work will examine antipsychotic polypharmacy in more detail, following recent successful development of algorithms to capture this.[24] As described earlier, utilising the keyword search functionality in CRIS, exposure to non-pharmacological agents such as khat was investigated,[14] and a large series of cases with suspected neuroleptic malignant syndrome were successfully identified which allowed a matched case–control study of antipsychotic exposures potentially responsible.[12 13] The association between antidepressant use and risk of mania and bipolar disorder has also recently been investigated,[50] as has antipsychotic use in children and adolescents with autistic spectrum disorder.[51] Finally, the potential to use extensive routine data to monitor treatment response was exemplified in a recent study of people receiving acetylcholinesterase inhibitor treatments for Alzheimer's disease in which trajectories of cognitive function were plotted before and after treatment initiation in order to identify predictors of 'response'—to our knowledge, the largest and most extensive cohort of its kind.[17]

Recent developments which are likely to generate substantial future output include the assembly of one of the largest cohorts to date of women with severe mental disorder who are followed from preconception and pregnancy to investigate medication use in relation to maternal and fetal outcomes.[52] Supplementing CRIS-derived outcomes to large clinical research samples with genetic profiling has also begun to generate novel output, for example, indicating that a well-recognised genetic risk factor for schizophrenia may also be a risk factor for worse clinical outcomes after diagnosis.[53] NLP applications have recently been extended to cover a range of affective and psychotic symptoms, allowing much more detailed phenotyping of large samples than a diagnosis alone provides,[54 55] and a range of adverse drug events have also recently been successfully captured.[56]

Table 3 Characteristics of active and inactive cases on the South London and Maudsley National Health Service (NHS) Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: most recent specialty (census date: 31 December 2014)*

Current or most recent SLaM specialty service providing care	Number (%)	
	Active patients	Inactive patients†
Psychosis	7116 (22.3)	12 444 (6.5)
Child and Adolescent Mental Health Services	5765 (18.0)	27 231 (14.2)
Mood, Anxiety and Personality	5271 (16.5)	31 887 (16.7)
Mental Health of Older Adults and Dementia	4217 (13.2)	24 842 (13.0)
Psychological Medicine	4333 (13.6)	59 212 (31.0)
Addictions	2559 (8.0)	12 768 (6.7)
Behavioural and Developmental Psychiatry	3532 (11.1)	7898 (4.1)
Unknown/not recorded	719 (2.2)	80 440 (42.1)

*Some patients may have records with more than one specialty.
†Inactive: those not currently receiving treatment and who have been discharged from all services.

Table 4 Characteristics of active and inactive cases on the South London and Maudsley National Health Service (NHS) Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: patterns of multispecialty care (census date: 31 December 2014)

	Number of specialties involved (current or most recent status)							
	Active patients			Inactive patients				
Specialty	1	2*	3+*	1	2*	3*	4*	5+*
Addictions	2349	197	13	9348	2181	903	315	21
Behavioural and Developmental Psychiatry	3347	178	<10	6484	962	324	114	14
Child and Adolescent Mental Health Services	5671	86	<10	24 299	2275	521	128	<10
Mental Health of Older Adults and Dementia	4173	42	<10	22 360	2159	261	55	<10
Mood, Anxiety and Personality	4653	582	36	19 595	8670	3105	493	24
Psychological Medicine	3818	481	34	40 778	14 199	3690	521	24
Psychosis	6509	580	27	4644	4522	2773	482	23
Total	30 520	1073	41	127 508	17 484	3859	527	24

*Include multiple counts of patients.

DISCUSSION

Currently, the SLaM BRC Case Register contains over 250 000 patient records and we believe it is the largest mental health data resource of its kind (ie, derived from the full EHRs for mental healthcare services). Since its original description, the database has nearly doubled in numbers of patients represented, but more importantly there have been key developments in the infrastructure to expand further the scale and depth of information available for research.[7] These developments have been primarily in NLP and linkage with external data sets.

Strengths and limitations of NLP

NLP is being applied increasingly to extract information from medical records, including applications for the detection of specific adverse drug events and other health events such as falls and nosocomial infections,[57–59] as well as use to identify obesity status and obesity-related diseases.[60 61] Furthermore, mining patient electronic medical records has been found to be useful for detecting patterns in patient care and patient treatment habits.[62 63] Statistical text mining has been used to determine if patients suffer from comorbidities

Table 5 Characteristics of active and inactive cases on the SLaM BRC Case Register: primary diagnoses ever recorded (census date: 31 December 2014)*

	Number (%)	
Assigned primary diagnosis (ICD-10 code and description)	Active patients	Inactive patients†
F0–F09—organic, including symptomatic, mental disorders	3517 (11.0)	19 535 (10.2)
F10–F19—mental and behavioural disorders due to psychoactive substance use	3742 (11.7)	19 204 (10.0)
F20–F29—schizophrenia, schizotypal and delusional disorders	6778 (21.2)	10 069 (5.3)
F30–F39—mood (affective) disorders	6076 (19.0)	31 119 (16.3)
F40–F48—neurotic, stress-related and somatoform disorders	4155 (13.0)	22 800 (11.9)
F50–F59—behavioural syndromes associated with physiological disturbances and physical factors	1025 (3.2)	5800 (3.0)
F60–F69—disorders of adult personality and behaviour	1518 (4.7)	4078 (2.1)
F70–F79—mental retardation	807 (2.5)	2050 (1.1)
F80–F89—disorders of psychological development	1483 (4.6)	4405 (2.3)
F90–F98—behavioural and emotional disorders with onset usually occurring in childhood and adolescence	3607 (11.3)	10 343 (5.4)
Unspecified mental disorder	7016 (22.0)	28 122 (14.7)
No axis 1 diagnosis	526 (1.6)	6399 (3.3)
G—diseases of the nervous system	173 (0.5)	543 (0.3)
Other illness codes (A–E, H–Q)	669 (2.1)	7292 (3.8)
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	101 (0.3)	1164 (0.6)
S–Y—injury, poisoning and external causes	398 (1.2)	1416 (0.7)
Z—factors influencing health status and contact with health services	6384 (20.0)	42 552 (22.2)
Number of patients with a primary diagnosis recorded (% of all patients)	29 820 (93.3)	157 027 (82.1)

*Some patients may have had more than one primary diagnosis recorded.
†Inactive: those not currently receiving treatment and who have been discharged from all services.
ICD, International Classification of Diseases; SLaM BRC, South London and Maudsley National Health Service (NHS) Foundation Trust Biomedical Research Centre.

Table 6 Number of patients represented on the SLaM BRC Case Register with CRIS data linked to other data sets

Data linkage	Number of patients on both databases (% of all CRIS active and inactive patients)
CRIS* and ONS mortality† data	20 864 (9.3)
CRIS* and HES‡ data	188 447 (84.4)
CRIS* and cancer registry§ data	3442 (1.5)

*CRIS active and inactive patients recorded as at 31 December 2014.
†(Up to 16 of December 2013.)
‡Up to 31 March 2013.
§(Cancer registry data last updated 31 December 2008).
CRIS, Clinical Record Interactive Search; HES, Health Episode Statistics; ONS, Office for National Statistics; SLaM BRC, South London and Maudsley National Health Service (NHS) Foundation Trust Biomedical Research Centre.

related to smoking, as well as detecting fall-related injuries, and regular expressions have been used to extract blood pressure values from progress notes.[64–66] NLP has been useful for extracting medical information such as principal diagnosis, information related to employment and medication use from clinical narratives.[64 67 68] This has led to a better understanding of the conditions patients face and potential interventions.[69] Manual chart review for annotation has been used extensively and when appropriate rigour is applied, the information extracted is very reliable and is often used as the reference standard to evaluate IE systems. Although the potential of NLP in mental health research was recognised in 1992, there have been few applications in clinical records from this specialty beyond those used for de-identification purposes.[70] However, progress is being made, including US studies using NLP to determine depression outcome, and adverse drug reactions, and characterisation of diagnostic profiles.[71–73]

Table 7 Number of people represented on the SLaM BRC Case Register with linked HES data

	CRIS data*	
HES data†	Active (%)	Inactive (%)
Any inpatient care‡	18 387 (57.5)	137 577 (71.9)
Any emergency room attendance‡	18 139 (56.8)	129 041 (67.5)
Any outpatient attendance‡	20 642 (64.6)	150 748 (78.8)

*CRIS active and inactive patients recorded as at 31 December 2014.
†Excluding mental health inpatient/outpatient services.
‡Excluding mental health providers.
CRIS, Clinical Record Interactive Search; HES, Health Episode Statistics; SLaM BRC, South London and Maudsley National Health Service (NHS) Foundation Trust Biomedical Research Centre.

Table 8 Number of people represented on the SLaM BRC Case Register with linked HES and mortality data

	Number of deaths (%)	
Data linkage sample	Total	Linked to ONS mortality records*
People in CRIS† with at least one inpatient admission in HES	20 541 (9.2)	19 910 (8.9)
People in CRIS† with at least one A&E attendance record in HES	14 791 (6.6)	14 279 (6.4)
People in CRIS† with at least one outpatient record in HES	19 220 (8.6)	18 613 (8.3)

*Up to 16 of December 2013.
†All CRIS active and inactive patient deaths recorded up to 16 December 2013.
A&E, accident and emergency; CRIS, Clinical Record Interactive Search; HES, Health Episode Statistics; ONS, Office for National Statistics; SLaM BRC, South London and Maudsley National Health Service (NHS) Foundation Trust Biomedical Research Centre.

Considering performances of NLP IE applications applied to clinical text, one study developed an NLP system for classifying patients with 15 comorbidity states for diseases related to obesity, found that the automated system performed well against manual expert rule-based systems, and concluded that even a relatively complex task was possible for an automated system on the basis of F-measures ranging from 0.48 for gastro-oesophageal reflux disease as a comorbidity to 0.96 for depression, and an overall system F-value of 0.60.[74] Another study evaluated automatic ascertainment of smoking status in 502 de-identified medical discharge records with 11 groups producing annotations and F-measures varying from 0.33 to 0.70 for current smoking status and 0.44 to 0.76 for past smoking.[75] F-measures for our applications were therefore relatively favourable. On the other hand, an application to identify and extract a patient's smoking status from clinical narrative text from Spanish outpatient records, evaluated against manual annotations, cited precision and recall statistics for a smoker versus non-smoker classification of 85% and 90%, respectively, and those for a current versus past smoker classification as 91% and 94%.[76] In our application, we achieved comparable precision but lower recall.[31]

Preliminary studies ascertaining postoperative complications using NLP have been cited as yielding encouraging results.[77 78] For example, in a recently conducted pilot study of statistical NLP for identifying cases of deep vein thrombosis (DVT) and pulmonary embolism (PE) from free-text electronic narrative radiology reports, the positive predictive value and sensitivity for DVT were 89% and 80%, respectively, and those for PE were 84% and 79%.[79] Another NLP application developed to ascertain weekly warfarin doses reported findings of

Table 9 Number of deaths in SLaM linked with ONS mortality data by underlying primary cause of death (latest date of record is as at 16 of December 2013)

ICD-10 chapter description (underlying cause of death)	Number of patients (% of all deaths in CRIS) (N=20 864)
Benign neoplasms or diseases of the blood	159 (0.8)
Cancers	3356 (16.1)
Certain conditions originating in the perinatal period and pregnancy, childbirth and the puerperium	<10
Codes for special purposes (eg, antibiotic resistance)	45 (0.2)
Congenital malformations, deformations and chromosomal abnormalities	73 (0.3)
Diseases of the circulatory system	5665 (27.2)
Diseases of the digestive system	1467 (7)
Diseases of the genitourinary system	689 (3.3)
Diseases of the musculoskeletal system	241 (1.2)
Diseases of the nervous system	1338 (6.4)
Diseases of the respiratory system	2964 (14.2)
Diseases of the skin	80 (0.4)
Endocrine, nutritional and metabolic diseases	445 (2.1)
External causes	1294 (6.2)
Infectious and parasitic diseases	356 (1.7)
Mental and behavioural disorders	2206 (10.6)
Symptoms and sign not elsewhere classified	376 (1.8)
Unknown/ missing	102 (0.5)

CRIS, Clinical Record Interactive Search; ICD, International Classification of Diseases; ONS, Office for National Statistics; SLaM, South London and Maudsley.

Table 10 Numbers of patients with both CRIS and cancer registry data, by primary cancer diagnosis (linkage last updated 31 December 2008)

Primary diagnosis (ICD-10 3-digit description)	Number (%) of patients
Malignant neoplasm of breast	563 (16.4)
Carcinoma in situ of cervix uteri	394 (11.4)
Malignant neoplasm of prostate	391 (11.4)
Malignant neoplasm of bronchus and lung	306 (8.9)
Malignant neoplasm of colon	179 (5.2)
Other malignant neoplasms of skin	152 (4.4)
Malignant neoplasm of bladder	92 (2.7)
Malignant neoplasm of rectum	90 (2.6)
Malignant neoplasm of corpus uteri	71 (2.1)
Malignant neoplasm of kidney, except renal pelvis	70 (2.0)
Other and unspecified types of non-Hodgkin's lymphoma	65 (1.9)
Malignant melanoma of skin	58 (1.7)
Malignant neoplasm of brain	57 (1.7)
Malignant neoplasm of pancreas	53 (1.5)
Malignant neoplasm of stomach	53 (1.5)
Malignant neoplasm without specification of site	53 (1.5)
Malignant neoplasm of oesophagus	50 (1.5)
Malignant neoplasm of cervix uteri	48 (1.4)
Malignant neoplasm of ovary	46 (1.3)
Diffuse non-Hodgkin's lymphoma	42 (1.2)
Myeloid leukaemia	42 (1.2)
Lymphoid leukaemia	39 (1.1)
Multiple myeloma and malignant plasma cell neoplasms	36 (1.0)
Malignant neoplasm of larynx	34 (1.0)
Other diagnoses	458 (13.3)

CRIS, Clinical Record Interactive Search; ICD, International Classification of Diseases.

90.8% precision and 99.7% recall, and a broader medication-ascertaining application achieved 86% precision and 77% recall.[68 80] In our own data, an evaluation of the NLP diagnosis application yielded a precision of 99% and a recall of 98% for vascular dementia, and our evaluations of the pharmacotherapy application found over 90% precision and recall for clozapine, although higher accuracy may be due to the combined use of structured data. It should be borne in mind that performances for one diagnosis or medication cannot be assumed to generalise to others, so it is still CRIS policy to advise de novo evaluation of application performance in studies investigating previously unevaluated entities. This is particularly pertinent to investigating antipsychotic medication prescribing, which is frequently preceded by clinical discussions and possibly tests (ie, clozapine); therefore, the presence of multiple annotations may not be reflective of current prescribing.

As displayed in table 12, the development of NLP IE applications to date has resulted in a very substantial expansion in data fields available for analysis within the SLaM BRC Case Register and in the ability to construct longitudinal data sets with repeated measures (as illustrated for MMSE score trajectories before and after initiation of dementia treatment).[17] With increasing use of EHRs, we believe that NLP techniques have an important role to play, whether derived metadata are to be used for research or to enhance the quality of the clinical record. This is particularly pertinent for mental health records where text fields are often substantial and contain some of the most important clinical information. However, although its potential is substantial, it is important to bear in mind that there may be limits in the usefulness of NLP in EHR-sourced data resources, because of the high degree of variability in clinical text. As well as the well-recognised challenges of non-grammatical sentences, misspellings, idiosyncratic abbreviations and jargon, there are more complex issues to deal with such as the establishment of temporality (eg, timing of events described in long case summaries), the classification of documents and within-document text domains (eg, sections of the history or mental state assessment), and the development of standard

Table 11 Performance of natural language processing information extraction applications developed to date in the SLaM BRC Case Register

Application name	Construct sought	Number of patients tested	Precision	Recall	F-statistic
Smoking[31]	Is the patient a current smoker?	100	0.93	0.58	0.72
Clozapine—current use[32]	Is the patient currently using clozapine (within 3 months)?	Precision: 279, recall: 200	0.96	0.92	0.94
Clozapine—ever used[32]	Has the patient used clozapine in the past?	Precision: 279, recall: 200	0.99	0.92	0.95
Diagnosis[33]	What text accompanies a statement about diagnosis?	75	0.99	0.98	0.99
MMSE[34]	What MMSE score did the patient attain on a given date?	100	0.97	0.98	0.97
Education	What age did a patient leave school?	Precision: 100, recall: 115	0.95	0.59	0.73
Living alone	Is the patient living alone?	100	0.93	0.99	0.96
Amisulpride[24]	Is the patient currently using amisulpride?	20 patients with 619 instances	0.97	0.61	0.75
Flupentixol[24]	Is the patient currently using flupentixol?	20 patients with 328 instances	0.94	0.77	0.85
Haloperidol[24]	Is the patient currently using haloperidol?	20 patients with 747 instances	0.94	0.57	0.71
Olanzapine[24]	Is the patient currently using olanzapine?	20 patients with 1150 instances	0.95	0.69	0.80
Risperidone[24]	Is the patient currently using risperidone?	20 patients with 737 instances	0.95	0.64	0.76
Zuclopenthixol[24]	Is the patient currently using zuclopenthixol?	20 patients with 390 instances	0.97	0.68	0.80

MMSE, Mini-Mental State Examination; SLaM BRC, South London and Maudsley National Health Service (NHS) Foundation Trust Biomedical Research Centre.

ontologies, not to mention the challenges of translation and harmonisation across languages. An important decision in NLP application development at the outset is whether near-perfect performance is required at an individual level, or whether a lower performance probabilistic approach might be appropriate. The latter may be sufficient for analyses to be carried out over large samples, but the former is likely to be required if the application is then to be used for clinical decision support.

Strengths and limitations of data linkages

As well as NLP applications, we were also able to expand the depth of information in this mental health case register through linkages with external data, including mortality, cancer and hospitalisation, with a primary care linkage recently developed and a linkage with education records fully approved and about to be implemented. Data linkage has been used in a variety of registers to enhance research questions. For example, nationwide data from the Icelandic Medicines Registry and the Database of National Scholastic Examinations were linked to study associations between drug treatment of attention deficit/hyperactivity disorder and academic performance.[81] In Sweden, acute myocardial infarction episodes were linked with routinely collected data on hospital discharges, mental health and mortality.[82] UK general practice data have been linked to national mortality, hospitalisation and disease register data at an individual level, and to census-derived socioeconomic data at a small area level.[83] The Western Australian e-cohort of half a million children included data cross-linked across a number of administrative registers including education, mental healthcare, hospital discharges, midwives notifications, cancer registrations, a registry of births, deaths and marriages and emergency presentations.[4]

Techniques for achieving both valid and secure data linkages within a robust governance framework are becoming increasingly standardised. In the Western Australian system, in order to protect privacy, linkage and analysis tasks are performed separately and linked data sets have identifiers removed before they are made available to researchers. Comparable procedures are followed in CRIS linkages. The data linkage process in Western Australia involves probabilistic methods to calculate the likelihood that two records belong to the same entity (person, family, event and location), whereas an important feature of the UK NHS is the NHS number, a unique reference for all patients, which we were able to use as the primary link for health-related information with CRIS data. Unique identifiers assigned at birth also exist in a number of other countries, including the unique citizen identifier, Civil Personal Registration number in Denmark covering prescription drug

Table 12 Summary of number of annotations generated from NLP applications in the SLaM BRC Case Register*

Application name	Total number of instances generated	Number of patients with at least one instance generated
MMSE	107 384	24 705
Diagnosis	615 237	78 851
Smoking	670 053	52 700
Education	181 905	51 665
Medication (selected)†		
Olanzapine	371 754	25 697
Citalopram	144 072	24 363
Mirtazapine	135 309	23 710
Risperidone	240 068	22 046
Zopiclone	129 488	20 712
Diazepam	129 409	17 841
Lorazepam	119 357	15 637
Fluoxetine	96 258	15 527
Sertraline	95 381	13 600
Promethazine	112 256	12 861
Clonazepam	111 279	9679
Quetiapine	98 509	9503
Aripiprazole	90 866	8737
Haloperidol	53 936	7591
Amisulpride	58 751	6759
Methadone	128 132	6385
Flupentixol	25 576	5248
Clozapine	111 170	4364
Zuclopenthixol	18 099	3093

*The CRIS database is updated every 24 h, so numbers are dynamic and displayed for illustrative purposes. NLP application run dates as follows: MMSE (24 June 2014), diagnosis (20 June 2014), smoking (17 July 2014), education (30 June 2014), medication (16 June 2014).
†Most frequent 15 agents plus those evaluated in table 11.
CRIS, Clinical Record Interactive Search; MMSE, Mini-Mental State Examination; NLP, natural language processing; SLaM BRC, South London and Maudsley National Health Service (NHS) Foundation Trust Biomedical Research Centre.

purchases, hospital inpatient, emergency and outpatient encounters, admissions to psychiatric hospitals, a range of disease-specific registries, primary care data and cause of death.[84] In Taiwan, social insurance enumeration systems have been used to create the National Health Insurance Research Database which has high national coverage and includes data from social insurance, health information, census and education resources.[85]

Record linkages are particularly valuable when they enable the capture of exposure data from one source and outcome data from another source, and have enabled novel investigations such as those attained through linking conscription surveys in Sweden and Israel with healthcare registers. Databases utilising the northern European system of unique citizen number will still have particular value in the following respects: (1) where information is gained on the total population within a geographic or administrative area, and not only insured patients; (2) where the person identifier is used for wider purposes than healthcare allowing novel and informative linkages, as discussed. The development of these linkages for the SLaM BRC Case Register is thus comparable with current practice elsewhere; however, the depth of information on mental healthcare accessed by CRIS is, we believe, currently unique in scale and scope, which we hope will enable findings from larger national samples to be further investigated in greater depth at a local level. There are various limitations with data linkage. First of all, most of the data linked to CRIS have time limitations, and cannot be used to develop decision support applications, because they are not available in real time. Mismatched identifier variables also place limits on the linkage process, although we have found this to be rare for the NHS number.

Collaborations

Work to date on the SLaM BRC Case Register has involved a number of welcomed collaborations, including those with other academic groups, both national and international, as well as with industry partners in pharmaceutical and biotech sectors. The authors particularly acknowledge the longstanding and fruitful collaboration with the University of Sheffield Department of Computer Science on the application of NLP techniques. The primary consideration with collaboration is the requirement (a component of the Case Register's ethics approval) that all data remain within the NHS firewall during analysis. In order to facilitate this, a dedicated office suite was set up in SLaM premises, the 'BRC Nucleus' to accommodate staff and visitors accessing Case Register data, although remote access, with appropriate security, is also possible. A second requirement is an appropriate affiliation with SLaM for those accessing the data, most usually taking the form of an honorary or substantive contract, or a 'research passport', but also covered on occasions by appropriate between-institution legal agreements as directed by the SLaM Caldicott Guardian—the statutory office overseeing the use of patient information in the NHS. All research projects using CRIS are considered and approved by a patient-led Oversight Committee, reporting to the Caldicott Guardian, as described in detail elsewhere.[8] As well as considering the appropriateness of research proposals, the CRIS Oversight Committee also adjudicate on risks of de-anonymisation at the analysis planning stage and, if needed, in the preparation of findings for publication (eg, proof-reading papers reporting quoted text excerpts).

Implications and challenges for future developments

Data derived from EHRs have huge potential to contribute to research and clinical care. Observational data are vital in healthcare-relevant research. As well as research into disease risk factors, incidence and prognosis, an important application of EHR-derived data is in providing 'real-world' information on response to routine clinical interventions (eg, recovery, adverse events) and, most importantly, predictors of response. The

ascertainment of characteristics predicting good/poor intervention response supports 'personalised medicine'. Compared with EHRs, randomised trials are insufficiently powered, even when combined, to detect predictors of response, and their samples are frequently highly selected—hence the need for large, generalisable data sets containing detailed information on routine clinical care. For example, the recently reported CRIS study of MMSE score trajectories before and after acetylcholinesterase inhibitor treatment initiation in dementia captured data on at least eight times more person-years of treatment from a single mental healthcare provider than all randomised controlled trial samples combined, as well as providing the added generalisability of 'real-world' data.[17] EHR databases also potentially allow enhanced and more effectively targeted recruitment for randomised controlled trials and other intervention evaluations, in addition to permitting pretrial modelling and efficiency planning. Approach for research study participation is generally considered to require prior consent (ie, 'opt in'), and a 'Consent for Contact' model for patient recruitment has been developed at SLaM.[86]

In the UK, EHRs are now near-ubiquitous in primary care and mental healthcare, and rapidly becoming so in acute care. However, realising their potential for clinical research depends heavily on the quality and nature of EHR data. In mental healthcare, applications have been very limited to date. In particular, although nearly all mental health services use EHRs, most clinically relevant information (eg, on symptoms, interventions, outcomes) is recorded in text and therefore not accessible for large-scale analyses to inform service planning, or for algorithms to support clinical decision-making. Given the very high individual and societal impact of disorders such as schizophrenia, bipolar disorder, depression and dementia, and the large mental healthcare sector, this data deficiency is a major limitation. For example, current national data on mental healthcare in the UK are principally available from three sources: (1) primary care data resources such as the Clinical Practice Research Datalink which covers approximately 5–10% of general practices;[87] (2) HES;[88] and (3) the Mental Health Minimum Data Set (MHMDS). However, each has key limitations. Primary care data do not contain information on mental health service interventions or sufficient information on the symptoms for which interventions are received and with which outcomes are evaluated. HES data are primarily used for identifying inpatient episodes and have limited data on interventions or outcomes beyond service receipt. The MHMDS covers mental healthcare more comprehensively; however, data are essentially restricted to service-level interventions (eg, pharmacotherapy is not recorded), and information on symptomatology and context for most patients is restricted to the relatively coarse Health of the Nation Outcome Scales.[89]

One solution for improving the structure of routine clinical data in the EHR would be to impose this structure at the point of data entry. However, the applicability of this approach depends on the willingness of clinical staff to input structured data; the accuracy of form completion; and on the extent to which the disorders, interventions and outcomes can be captured in pre-prepared scales. Our experience has been that imposition of structured fields in a clinical record is difficult to achieve, and even more so to sustain, at least within mental healthcare. Furthermore, although a structured field improves data accessibility, it does not necessarily render the data any more valid. Even in a clinical context where data have inherent structure (eg, blood pressure recordings following hypertension treatment), this approach has limitations and may fail to capture influential contextual factors (eg, suboptimal adherence to antihypertensive treatment, or 'white coat hypertension'). Application of structure is particularly challenging in mental healthcare where interventions are primarily determined by qualitatively reported experiences (symptoms), where outcomes rely on tracking improvement or deterioration of the same constructs, and where some interventions themselves are not readily prestructured (eg, psychotherapeutic strategies). Although constructs such as medication sound amenable to imposed structure, this is limited in UK services because of the mixed prescribing between primary and mental healthcare. Structured recording of current medication outside a prescribing database is difficult to maintain with any accuracy because there is no clear gain for clinicians to enter medication receipt in a structured field compared with recording the same information in text. We have demonstrated that it is feasible to obtain at least some novel structured information from routine mental health records on a range of clinical indicators using NLP. The over-riding advantage of this approach is that no additional 'data entry' is required by clinical staff beyond what is normal practice. The validity of the approach has been demonstrated in a typical mental health service EHR at SLaM and it is reasonable to suppose at least some generalisability to other UK mental health services, given the relatively standardised nature of clinical assessments and national training in psychiatry. However, clearly cross-applicability is important to evaluate and in this respect it is advantageous that the CRIS application was successfully implemented in 2014 at four other mental health Trusts with comparable EHR systems (http://www.slam.nhs.uk/research/d-cris). Finally, as with all data derived from routine sources, it is important to bear in mind, when designing investigations, the reasons why information may or may not be recorded in clinical practice—including the incentives for recording within different clinical services or at different points on the healthcare pathway. For example, in early analyses using the application to ascertain current smoking status, it was found that missing data were relatively high unless the focus was on

patients who had received at least a year's care from SLaM.[31] Enhancing the structure of a record could be one answer, although better design and focusing of text fields may in the end be more acceptable.

A more generic challenge for the use of specialist healthcare data lies in the limited time 'windows' within which data are provided. Cohort studies using such data resources therefore need to take into account not only what data are available from the record but also the time periods within which they are available. These time periods also need to be carefully considered in relation to the question under investigation, since they are determined by discharge and/or re-referral, which clearly themselves are determined by factors such as recovery, engagement with services and out-migration from the catchment. Those patients on whom longest periods of follow-up are available are likely to be those who have more severe symptomatology (requiring longer periods of care), although they may also have more stable accommodation or support and thus less likelihood of out-migration. Data linkages can provide some means of addressing the problem—for example, national data on hospitalisation or mortality accrue regardless of a patient's contact or not with mental healthcare; however, these may be limited in depth of information, as described above.

A key challenge inherent with all use of healthcare data is how to ensure such data are appropriately and robustly protected and how to develop and to use anonymised clinical information in a way that is acceptable to the general public, and most importantly to patients. Such challenges incorporate not only a case register's data themselves but also procedures around data linkage where use of identifiers is required, although systems are increasingly becoming established which achieve data linkage in ways that effectively preserve anonymity. Data protection laws and practice vary internationally, but most do have some provision for the use of data without prior consent if these data are effectively anonymised and if important research cannot be carried out in any other way. It is also worth bearing in mind at the outset that few data sets can be claimed to be wholly anonymised. For example, even in the shallowest of administrative databases, a combination of age, gender and date/place of admission might well be sufficiently unique that it theoretically identifies a person. Technical solutions to anonymisation are therefore never sufficient on their own, but need to be accompanied by a governance structure which evaluates database use for any risk of compromising anonymity, as well as monitoring the appropriateness of the research being carried out, and of the people and agencies having data access. The coming years will bring many more opportunities for the use and linking of anonymised EHR data. It is clear that researchers, patients and the general public need to be engaged in ongoing conversations and collaborations to develop appropriate frameworks so as to maximise the use of such data in ways that maintain the trust of all parties. The SLaM BRC Case Register involved patients from the outset both in designing the security model and in leading ongoing oversight of data use and dissemination,[8] thus ensuring that discussions about the future of EHR use (scientifically, and as a sociological question) effectively and meaningfully engage the stakeholders whose data have generated the resource in the first place.

Author affiliations
[1]King's College London (Institute of Psychiatry, Psychology and Neuroscience), London, UK
[2]South London and Maudsley NHS Foundation Trust, London, UK
[3]Durham University, Durham, UK

Contributors The cohort is led by RS and MB who conceived the study and manuscript. The cohort description was led by GP and AT. All named authors initially contributed significant text to the cohort description. Analyses were carried out by MB, C-KC, RDH, GK, RL and HS. Descriptions of the database were led by MB, AF, AJ, MH and MP. Descriptions of data linkages were led by RL, JD, RD and MH. Descriptions of text use were led by RJ, RDH, GK and AT. FC contributed on governance and oversight. All authors reviewed, contributed to and approved the final manuscript.

Funding RD is funded by a Clinician Scientist Fellowship from the Health Foundation in partnership with the Academy of Medical Sciences. RDH is funded by a Medical Research Council (MRC) Population Health Scientist Fellowship (grant number MR/J01219 X/1). FC's research is supported by the Wellcome Trust. The data resource and all other authors are funded by the National Institute for Health Research (NIHR) Biomedical Research Centre and Dementia Biomedical Research Unit at South London and Maudsley NHS Foundation Trust and King's College London.

Disclaimer The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Competing interests RDH, C-KC, RJ, HS, MB and RS have received research funding from Roche, Pfizer, J&J and Lundbeck.

Ethics approval Oxford REC C.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. Ten Horn HMM, Gie LR, Gulbinat WH, *et al. Psychiatric case registers in public health. A worldwide inventory 1960–1985.* Amsterdam: Elsevier, 1986.
2. Perera G, Soremekun M, Breen G, *et al.* The psychiatric case register: noble past, challenging present, but exciting future. *Br J Psychiatry* 2009;195:191–3.
3. Allebeck P. The use of population based registers in psychiatric research. *Acta Psychiatr Scand* 2009;120:386–91.
4. Morgan VA, Assen V, Jablensky AV. From inventory to benchmark: quality of psychiatric case registers in research. *Br J Psychiatry* 2010;197:8–10.
5. Stewart R. The big case register. *Acta Psychiatr Scand* 2014;130:83–6.
6. Amaddeo F. The small scale clinical psychiatric case registers. *Acta Psychiatr Scand* 2014;130:80–2.
7. Stewart R, Soremekun M, Perera G, *et al.* The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) case register: development and descriptive data. *BMC Psychiatry* 2009;9:51.
8. Fernandes AC, Cloete D, Broadbent MT, *et al* Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. *BMC Med Inform Decis Making* 2013;11:71.

9. World Health Organisation, 1983. *Psychiatric case registers. Report on a Working Group*. Copenhagen: WHO Regional Office for Europe.
10. Backus LI, Gavrilov S, Loomis TP, *et al*. Clinical Case Registries: simultaneous local and national disease registries for population quality management. *J Am Med Inform Assoc* 2009;16: 775–83.
11. Morden NE, Berke EM, Welsh DE, *et al*. Quality of care for cardiometabolic disease: associations with mental disorder and rurality. *Med Care* 2010;48:72–8.
12. Chang CK, Harrison S, Lee W, *et al*. Ascertaining instances of neuroleptic malignant syndrome in a secondary mental health care electronic medical records database: the SLAM BRC Case Register. *Ther Adv Psychopharmacol* 2012;2:75–83.
13. Su YP, Chang CK, Hayes RD, *et al*. Retrospective chart review on exposure to psychotropic medications associated with neuroleptic malignant syndrome. *Acta Psychiatr Scand* 2014;130:52–60.
14. Tulloch AD, Frayn E, Craig TK, *et al*. Khat use among Somali mental health service users in South London. *Soc Psychiatry Psychiatr Epidemiol* 2012;47:1649–56.
15. Tulloch AD, Fearon P, David AS. Residential mobility among patients admitted to acute psychiatric wards. *Health Place* 2011;17:859–66.
16. Tulloch AD, Fearon P, David AS. Timing, prevalence, determinants and outcomes of homelessness among patients admitted to acute psychiatric wards. *Soc Psychiatry Psychiatr Epidemiol* 2012;47:1181–91.
17. Perera G, Khondoker M, Broadbent M, *et al*. Factors associated with response to acetylcholinesterase inhibition in dementia: a cohort study from a secondary mental health care case register in London. *PLoS ONE* 2014;9:e109484.
18. Cunningham H. Information extraction, automatic. In: Brown K, ed. *Encyclopedia of language and linguistics*. 2nd edn. Elsevier, 2005:665–77.
19. Cunningham H, Maynard D, Bontcheva K, *et al*. GATE: a framework and graphical development environment for robust NLP tools and applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*; Philadelphia, 2002.
20. Cunningham H. GATE, a general architecture for text engineering. *Comput Humanit* 2002;36:223–54.
21. GATE research projects. http://gate.ac.uk/projects.html
22. Sager N, Lyman M, Bucknall C, *et al*. Natural language processing and the representation of clinical data. *J Am Med Inform Assoc* 1994;1:142–60.
23. Meystre SM, Savova GK, Kipper-Schuler KC, *et al*. Extracting information from textual documents in the electronic health record: a review of recent research. *Year Med Inform* 2008:128–44.
24. Kadra G, Stewart R, Shetty H, *et al*. Extracting antipsychotic polypharmacy data from electronic health records: developing and evaluating a novel process. *BMC Psychiatry* 2015;15:166.
25. Gorrell G, Jackson R, Roberts A, *et al*. Finding negative symptoms of schizophrenia in patient records. *Proc NLP Med Biol Work (NLPMedBio), Recent Adv Nat Lang Process (RANLP)*; 2013: 9–17.
26. Jackson R, Stewart R, Patel R, *et al*. TextHunter—a user friendly tool for extracting generic concepts from free text in clinical research. *Proc Am Med Inform Assoc* 2014;19:729–38.
27. Woodhead C, Ashworth M, Schofield P, *et al*. Patterns of physical co-/multi-morbidity among patients with serious mental illness: a London borough-based cross-sectional study. *BMC Fam Pract* 2014;15:117.
28. Schofield P, Baawuah F, Seed PT, *et al*. Managing hypertension in general practice: a cross-sectional study of treatment and ethnicity. *Br J Gen Pract* 2012;62:e703–9.
29. Chang CK, Hayes RD, Broadbent MT, *et al*. A cohort study on mental disorders, stage of cancer at diagnosis and subsequent survival. *BMJ Open* 2014;4:e004295.
30. Office for National Statistics, 2011. http://www.neighbourhood.statistics.gov.uk/dissemination/Download1.do?&nsjs=true&nsck=false&nssvg=false&nswid=1600 (accessed 8 Aug 2014).
31. Wu CY, Chang CK, Robson D, *et al*. Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register. *PLoS ONE* 2013;8:e74262.
32. Hayes RD, Downs J, Chang CK, *et al*. The effect of clozapine on premature mortality: an assessment of clinical monitoring and other potential confounders. *Schizophr Bull* 2015;41:644–55.
33. Sultana J, Chang CK, Hayes RD, *et al*. Associations between risk of mortality and atypical antipsychotic use in vascular dementia: a clinical cohort study. *Int J Geriatr Psychiatry* 2014;29:1249–54.
34. Su YP, Chang CK, Hayes RD, *et al*. Mini-mental state examination as a predictor of mortality among older people referred to secondary mental healthcare. *PLoS ONE* 2014;9:e105312.
35. Williams P, Csipke E, Rose D, *et al*. Efficacy of a triage system to reduce length of hospital stay. *Br J Psychiatry* 2014;204:480–5.
36. Brown PF, Tulloch AD, Mackenzie C, *et al*. Assessments of mental capacity in psychiatric inpatients: a retrospective cohort study. *BMC Psychiatry* 2013;13:115.
37. Patel R, Shetty H, Jackson R, *et al*. Delays before diagnosis and initiation of treatment in patients presenting to mental health services with bipolar disorder. *PLoS ONE* 2015;10:e0126530.
38. Fusar-Poli P, Diaz-Caneja CM, Patel R, *et al*. Services for people at high risk improve outcomes in patients with first episode psychosis. *Acta Psychiatr Scand* 2016;133:76–85.
39. Oram S, Khondoker M, Abas M, *et al*. Characteristics of trafficked adults and children with severe mental illness: a historical cohort study. *Lancet Psychiatry* 2015;2:1084–91.
40. Chang CK, Hayes RD, Broadbent M, *et al*. All-cause mortality among people with serious mental illness (SMI), substance use disorders and depressive disorders in southeast London: a cohort study. *BMC Psychiatry* 2010;10:77.
41. Hayes RD, Chang CK, Fernandes A, *et al*. Associations between substance use disorder sub-groups, life expectancy and all-cause mortality in a large British specialist mental healthcare service. *Drug Alcohol Depend* 2011;118:56–61.
42. Chang CK, Hayes RD, Perera G, *et al*. Life expectancy at birth for people with serious mental illness, substance use disorders, and depressive disorders from a secondary mental health care case register in London. *PLoS ONE* 2011;6:e19590.
43. Bogdanowicz KM, Stewart R, Broadbent M, *et al*. Double trouble: psychiatric comorbidity and opioid addiction-All-cause and cause-specific mortality. *Drug Alcohol Depend* 2015;148:85–92.
44. Hayes RD, Chang CK, Fernandes AC, *et al*. Functional status and all-cause mortality in serious mental illness. *PLoS ONE* 2012;7: e44613.
45. Hayes RD, Chang CK, Fernandes A, *et al*. Associations between symptoms and all-cause mortality in individuals with serious mental illness. *J Psychosom Res* 2012;72:114–19.
46. Wu CY, Chang CK, Hayes RD, *et al*. Clinical risk assessment rating and all-cause mortality in secondary mental healthcare: the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) Case Register. *Psychol Med* 2012;42:1581–90.
47. Fok ML, Stewart R, Hayes RD, *et al*. Predictors of natural and unnatural mortality among patients with personality disorder: evidence from a large UK case register. *PLoS ONE* 2014;9: e100979.
48. Fok ML, Hayes RD, Chang CK, *et al*. Life expectancy at birth and all-cause mortality among people with personality disorder. *J Psychosom Res* 2012;73:104–7.
49. Ward G, Perera G, Stewart R. Predictors of mortality for people aged over 65 years receiving mental healthcare for delirium, in a South London Mental Health Trust, UK: a retrospective survival analysis. *Int J Geriatr Psychiatry* 2015;30:639–46.
50. Patel R, Reiss P, Shetty H, *et al*. Do antidepressants increase the risk of mania and bipolar disorder in people with depression? A retrospective electronic case register cohort study. *BMJ Open* 2015;5:e008341.
51. Downs J, Hotopf M, Ford T, *et al*. Clinical predictors of antipsychotic use in children and adolescents with autism spectrum disorders: a historical open cohort study using electronic health records. *Eur Child Adolesc Psychiatry* 2015. Oct 15 [Epub ahead of print].
52. Taylor CL, Stewart R, Ogden J, *et al*. The characteristics and health needs of pregnant women with schizophrenia compared with bipolar disorder and affective psychosis. *BMC Psychiatry* 2015;15:88.
53. Wickramasinghe A, Tulloch AD, Hayes RD, *et al*. Associations between the schizophrenia susceptibility gene ZNF804A and clinical outcomes in psychosis. *Transl Psychiatry* 2015;5:e698.
54. Patel R, Lloyd T, Jackson R, *et al*. Mood instability is a common feature of mental health disorders and is associated with poor clinical outcome. *BMJ Open* 2015;5:e007504.
55. Patel R, Jayatilleke N, Broadbent M, *et al*. Negative symptoms in schizophrenia: a study in a large clinical sample of patients using a novel automated method. *BMJ Open* 2015;5:e007619.
56. Iqbal E, Mallah R, Jackson RG, *et al*. Identification of adverse drug events from free text electronic patient records and information in a large mental health case register. *PLoS ONE* 2015;10:e0134208.
57. Chazard E, Ficheur G, Merlin B, *et al*. PSIP consortium, Beuscart R. Detection of adverse drug events detection: data aggregation and data mining. *Stud Health Technol Inform* 2009;148:75–84.
58. Bates DW, Evans RS, Murff H, *et al*. Detecting adverse events using information technology. *J Am Med Inform Assoc* 2003;10:115–28.

59. Mendonca EA, Haas J, Shagina L, *et al*. Extracting information on pneumonia in infants using natural language processing of radiology reports. *J Biomed Inform* 2005;38:314–21.
60. Yang H, Spasic I, Keane JA, *et al*. A text mining approach to the prediction of disease status from clinical discharge summaries. *J Am Med Inform Assoc* 2009;16:596–600.
61. Guillen R. Identifying obesity and co-morbidities from medical records. *Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*; 2009:868.
62. Pakhomov SV, Hanson PL, Bjornsen SS, *et al*. Automatic classification of foot examination findings using clinical notes and Machine learning. *J Am Med Inform Assoc* 2008;15:198–202.
63. Rao RB, Krishnan S, Niculescu RS. Data mining for improved cardiac care. *SIGKDD Explor News* 2006;8:3–10.
64. Zeng QT, Goryachev S, Weiss S, *et al*. Extracting principal diagnosis, comorbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Med Inform Decis Mak* 2006;6:30.
65. Chiarini-Tremblay M, Berndt DJ, Foulis P, *et al*. Utilizing text mining techniques to identify fall related injuries. *Inf Technol Manag* 2009;10:226–53.
66. Turchin A, Kolatkar NS, Grant RW, *et al*. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. *J Am Med Inform Assoc* 2006;13:691–5.
67. Dillahunt-Aspillaga C, Finch D, Massengale J, *et al*. Using information from the electronic health record to improve measurement of unemployment in service members and veterans with mTBI and post-deployment stress. *PLoS ONE* 2014;9:e115873.
68. Xu H, Jiang M, Oetjens M, *et al*. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. *J Am Med Inform Assoc* 2011;18:387–91.
69. Cerrito P, Cerrito J. Data and text mining the electronic medical record to improve care and to lower costs. *Proceedings of the 31st Annual SAS Users Group International Conference*; 26–29 March 2006, San Francisco, CA.
70. Garfield DA, Rapp C, Evens M. Natural language processing in psychiatry. Artificial intelligence technology and psychopathology. *J Nerv Ment Dis* 1992;180:227–37.
71. Perlis RH, Iosifescu DV, Castro VM, *et al*. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychol Med* 2012;42:41–50.
72. Sohn S, Kocher JP, Chute CG, *et al*. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *J Am Med Inform Assoc* 2011;18(Suppl 1):i144–9.
73. Roque FS, Jensen PB, Schmock H, *et al*. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput Biol* 2013;7:e1002141.
74. Ambert KH, Cohen AM. A system for classifying disease comorbidity status from medical discharge summaries using automated hotspot and negated concept detection. *J Am Med Inform Assoc* 2009;16:590–5.
75. Uzuner O, Goldstein I, Luo Y, *et al*. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc* 2008;15:14–24.
76. Figueroa RL, Soto DA, Pino EJ. Identifying and extracting patient smoking status information from clinical narrative texts in Spanish. *Conf Proc IEEE Eng Med Biol Soc* 2014;2014:2710–13.
77. Murff HJ, FitzHenry F, Matheny ME, *et al*. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA* 2011;306:848–55.
78. FitzHenry F, Murff HJ, Matheny ME, *et al*. Exploring the frontier of electronic health record surveillance: the case of postoperative complications. *Med Care* 2013;51:509–16.
79. Rochefort CM, Verma AD, Eguale T, *et al*. A novel method of adverse event detection can accurately identify venous thromboembolism (VTEs) from narrative electronic health record data. *J Am Med Inform Assoc* 2015;22:155–65.
80. Spasic I, Sarafraz F, Keane JA, *et al*. Medication information extraction with linguistic pattern matching and semantic rules. *J Am Med Inform Assoc* 2010;17:532–5.
81. Zoëga H, Rothman KJ, Huybrechts KF, *et al*. A population-based study of stimulant drug treatment of ADHD and academic progress in children. *Pediatrics* 2012;130:e53–62.
82. Hammar N, Alfredsson L, Rosén M, *et al*. A national record linkage to study acute myocardial infarction incidence and case fatality in Sweden. *Int J Epidemiol* 2001;30(Suppl 1):S30–4.
83. Gerber DR, Bekes CE, Parrillo JE. Economics of critical care: Medicare part A versus part B payments. *Crit Care Med* 2006;34 (Suppl):S82–7.
84. Helweg-Larsen K, Kruse M. Violence against women and consequent health problems: a register-based study. *Scand J Public Health* 2003;31:51–7.
85. Chen VC, Wang TN, Liao YT, *et al*. Asthma and self-harm: a population-based cohort study in Taiwan. *J Psychosom Res* 2014;77:462–7.
86. Callard F, Broadbent M, Denis M, *et al*. Developing a new model for patient recruitment in mental health services: a cohort study using Electronic Health Records. *BMJ Open* 2014;4:e005654.
87. Williams T, van Staa T, Puri S, *et al*. Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. *Ther Adv Drug Saf* 2012;3:89–99.
88. Keown P, Mercer G, Scott J. Retrospective analysis of hospital episode statistics, involuntary admissions under the Mental Health Act 1983, and number of psychiatric beds in England 1996–2006. *BMJ* 2008;337:a1837.
89. Meagher D, O'Brien S, Pullela A, *et al*. Multidisciplinary activities in a community mental health service: relationship to the Health of the Nation Outcome Scales scores and diagnosis. *Psychiatric Bulletin* 2009;33:172–5.

BMJ Open

Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource

Gayan Perera, Matthew Broadbent, Felicity Callard, Chin-Kuo Chang, Johnny Downs, Rina Dutta, Andrea Fernandes, Richard D Hayes, Max Henderson, Richard Jackson, Amelia Jewell, Giouliana Kadra, Ryan Little, Megan Pritchard, Hitesh Shetty, Alex Tulloch and Robert Stewart

BMJ Open 2016 6:
doi: 10.1136/bmjopen-2015-008721